Generalized Linear Model-based Control Charts for High-Purity Processes

An exploratory study on the performance of GLM-based control charts with an application in the chemical industry

Gommers, M.E.

Award date:
2021

Department of Mathematics & Computer Science

Chemical Company Dow Inc.

Master's thesis

Eindhoven, August 2021

Generalized Linear Model-based Control Charts for High-Purity Processes

An exploratory study on the performance of GLM-based control charts with an application in the chemical industry

Supervisors:

At TU/e
Dr. A. Di Bucchianico
Email: a.d.bucchianico@tue.nl

Author:

M.E. Gommers
ID: 0904538, Email:
m.e.gommers@student.tue.nl

At Dow
C. Rizzo
Email: crizzo@dow.com

Abstract

Technology-enhanced solutions drive to continuous improvement of production processes in the chemical industry. An increasing number of manufacturing systems is equipped with digital devices that continuously monitor the amount of nonconforming items as indication of process quality and efficiency. Quality control of highly efficient production processes becomes a challenging task when produced items are mostly zero-defect. An additional complication arises when some covariate affects the amount of nonconforming items. At Dow Inc., plastic pellets are manufactured and monitored for defects, while the defect rate depends on a variable inspected weight.

In this Master's thesis, we consider generalized linear model-based control charts for detecting contextual anomalies in data that originates from monitoring high-purity processes for defects. Observations are assumed to follow a zero-inflated Poisson (ZIP) or zero-inflated negative binomial (ZINB) distribution that depend on a normally distributed covariate. The ZIP and ZINB regression models are employed for monitoring predictive Pearson, deviance and randomised quantile residuals in a regression-based Shewhart chart with both symmetric and probability control limits. A simulation study is proposed to compare the performance of each monitoring scheme. Results show that both the ZIP and ZINB regression-based Shewhart charts with deviance residuals and probability control limits perform satisfactory.

In addition, a Gamma GLM-based time-between-events chart is introduced for detecting contextual anomalies in high-purity count data. Simulation results show that the Gamma GLM-based TBE charts perform equally well with Pearson, deviance and quantile residuals.

Keywords: Statistical process control, high-purity processes, univariate monitoring, contextual anomaly detection, zero-inflated Poisson regression, zero-inflated negative binomial regression, Pearson residuals, deviance residuals, quantile residuals.

Preface

This Master's thesis has been written as a part of my final project for the Master Industrial and Applied Mathematics at the Eindhoven University of Technology. The project has been part of a graduation internship at Dow Chemical Company in Terneuzen, the Netherlands. Before continuing to the research, I want to take a moment to thank everyone who has made this thesis possible.

First of all, I want to thank Alessandro Di Bucchianico and Caterina Rizzo for the guidance that they gave me throughout this project. You introduced me to the field of statistical process control, which has been a pleasure to discover. Your feedback and support has guided me at times when the project was challenging, which I believe has helped me grow as a mathematician. You have thought me to stay critical at any time, and to persevere when things get complicated.

I also want to thank all the people at Dow, from the DF-FPS-CI group in Terneuzen, as well as the people from the global ChemometricsAI & Statistics group that I have worked with over the last six months. Even though it was a strange time and physical distance was required, you have all put in great effort to make me feel welcome. This is truly appreciated. I want to thank May Roca and Swee-Teng Chin especially, for being continuously involved in my work and for always showing interest in my well-being during the internship.

Last but not least, I want to thank my friends and family. I want to thank mom and dad for supporting me on every path I go. You have raised me with the confidence to make my own decisions, and it is thanks to you that I am able to write this thesis in the first place. To Rogier, Daan en Floor I want to say - thank you for always keeping me happy and sane. Finally, I want to thank my friends Alissa, Bauke, Bouke, Carmen, Daan, Ellen, Laura, Martijn, Merel, Niels, Sven and Puck. We have been friends since the first weeks of college and you have made my student time unforgettable.

Marieke Gommers, August 2021

Contents

List of Figures

List of Tables

Glossary

ACF	Autocorrelation function
ALI	Average length of inspection
ANOS	Average number of observations to signal
ARL	Average run length
BFGS	Broyden–Fletcher–Goldfarb–Shanno
CCC	Cumulative count control (chart)
CI	Continuous improvement
CL	Centre line
COM	Conway-Maxwell
CRL	Conforming run length
CUSUM	Cumulative sum
EDM	Exponential dispersion model
EM	Expectation-maximisation
EWMA	Exponentially weighted moving average
GLM	Generalized linear model
IC	In control
IWLS	Iterative weighted least squares
LCL	Lower control limit
LI	Length of inspection
LRT	Likelihood-ratio test
ML	Maximum likelihood
OC	Out-of-control
OLS	Ordinary least squares
PCA	Principal component analysis
RL	Run length
SDRL	Standard deviation of the run length
SPC	Statistical process control
SSE	Sum of squared estimate of errors
TBE	Time-between-events
UCL	Upper control limit
ZI	Zero-inflated
ZINB	Zero-inflated negative binomial
ZIP	Zero-inflated Poisson

1 Introduction

In our everyday lives, we often talk about quality. We talk about quality of goods and services, but we also talk about the quality of systems and processes. We generally seek the products that have high quality and we aim to improve the quality of those things that we offer to others. Quality improvement of products and processes generally leads to efficient use of materials and more sustainable industries. Over the years, this has led to a growing interest in the tools and methodologies that regard quality control. But what is quality? And how do we achieve it?

The 21st century answer to this question is: data science. We live in a world where data becomes more available every day and we are rapidly developing the skills to transform this data into valuable information. Many scientists believe that we are now at the beginning of the fourth industrial revolution, in which the combination of technology-enhanced solutions with information technology and data analytics will completely change the way in which we operate. Continuous optimisation of processes will significantly improve productivity and quality across all sectors such as finance, health care, marketing and manufacturing. In this Master's thesis, we focus on data science applications for quality control in the manufacturing sector. More specifically, we consider generalized linear model-based control charts for high-purity processes. This study is conducted as part of a research collaboration between Eindhoven University of Technology and the chemical company Dow Inc.

1.1 Dow

Dow was founded in 1897 and has grown to be one of the words largest chemical manufacturers in the world. The production site in Terneuzen, The Netherlands, counts 17 factories and is one of the largest production location of Dow worldwide. The site in Terneuzen hosts many facilities including production of goods, R&D and supporting functions for all sites in Europe, the Middle-East, Africa and India. The research of this Master's thesis is commissioned by the Chemometrics&AI and Statistics group in the Continuous Improvement (CI) organisation. The Chem&AI and Stats group strives to support the manufacturing and engineering organisations with data-driven solution for both short-term problems and long-term strategies. A strategic project is the implementation and support of monitoring systems for quality and process manufacturing data, providing plant personnel with real-time information about the state of the processes. Monitoring production processes has led to better decision making so far. However, it is known that standard monitoring strategies are not always appropriate for complex systems, particularly for those referred to as high-purity processes. The work of this thesis aims to contribute in building a more advanced monitoring systems framework in the context of high-purity processes.

1.2 Monitoring high-purity processes

Monitoring production output is a conventional method to obtain real-time information regarding production process performances. All produced goods at Dow are compared to their design requirements, which are called specifications, at the end of the production line. A product that does not meet the requirements

is labelled as 'defective' while all conforming products are labelled as 'non-defective'. The total amount of defective products per inspected sample is an indication of the production process performance over time, whereas fewer defects indicate better performance. If all products in the sample are non-defective, then this sample is referred to as having zero defect. Highly efficient production processes of which the monitoring data reports mostly zero defect are in the chemical industry referred to as high-purity process, and in other fields as high-yield processes. In this research project we consider the high-purity processes at Dow, and the way in which we can control their quality.

Statistical process control (SPC) is a field of research that considers statistical tools and methods for monitoring the quality of industrial processes over time. The textbook of Qiu (2013) provides an overview of the most important concepts and tools in SPC, of which a more elaborate description is provided in Chapter 4. SPC control charts monitor a quality characteristic over time, to determine whether the production process is still performing satisfactory. The amount of defective products per sample is for example a possible quality characteristic. An important assumption of SPC control charts is all observations are independent and identically distributed. However, this is not always true in practice.

It might be that the outcome of each observation depends on other variables as well. In such cases, it is necessary to account for the effect that these covariates have on the outcome, when monitoring these observations over time. Hence, we should not aim to detect abnormalities in the data, but we should detect events that are abnormal in their context. These events are called contextual anomalies. A method for detecting contextual anomalies is the application of regression-based control charts. We distinguish between traditional control charts and regression-based control charts, whereas traditional control charts aim to detect abnormalities in the data itself. Regression-based control charts aim to detect data points that are abnormal in their context, i.e., contextual anomalies.

At Dow, plant personnel use traditional control charts to monitor the health of the processes in all sites around the world. However, constructing regression-based control charts for high-purity processes can be complex due to complicated data distributions with excessive amounts of zero-occurrences. Nevertheless, the demand for such monitoring schemes remains. Therefore, in this thesis we focus on the application of regression-based control charts for high-purity processes, in order to provide Dow with a general framework to detect contextual shifts.

1.3 Research questions

It is common knowledge in SPC that generally there is not one monitoring strategy that is superior to all others. An optimal scheme should rather be designed according to a specific process and its monitoring goal. The goal of this thesis is therefore to provide a framework for identifying a monitoring scheme, that performs satisfactory given a specific high-purity process and monitoring goal. The main research question is defined as follows: Which monitoring scheme is most appropriate for detecting contextual anomalies in univariate count data, that originates from monitoring a specific high-purity processes? Sub-questions that arise in this setting are:

- Based on published works in the literature, what are the established monitoring methods for detecting contextual anomalies in data that originates from monitoring high-purity processes for defects?
- How can we model the relationship between the response variable and the covariate? And what type of residuals can we use for a regression-based control chart?
- Which regression-based monitoring schemes can be used for detecting contextual anomalies in data that originates from monitoring high-purity processes for defects?
- How can we evaluate the performance of a regression-based control chart?
- Which monitoring scheme achieves the best performance when aiming to detect contextual anomalies in data that originates from monitoring high-purity processes for defects?

1.4 Thesis outline

Before we proceed to answer any of the research questions, it is important become acquainted with the data at hand. Hence, a description of the high-purity monitoring data is provided in Chapter 2, along with a definition of two distributions that can be used to model this data. Next, a literature review on regression-based control charts for high-purity processes is presented in Chapter 3, where the solution strategy of this project is also provided. A certain level of understanding in SPC theory is assumed in the literature review, such that inexperienced readers might want to start with Chapter 4. Here, the main concepts of SPC and SPC control charts are discussed.

A traditional control chart is proposed in Chapter 5, where we focus on monitoring high-purity count data that is not affected by a covariate. We continue by considering data that is affected by a covariate, such that regression models for count data are discussed in Chapter 6. Monitoring methods for detecting contextual anomalies in high-purity count data are described Chapter 7, along with two strategies for performance evaluation of each method. The performance results are provided and discussed in Chapter 8. Finally, a new monitoring method is introduced as a suggestion for future work in Chapter 9, where we consider a regression-based time-between-events chart. Conclusions and discussion of this project are provided in Chapter 10. The technical mathematical background is attached in Appendix A, and additional results can be found in Appendix B. The `R` code of the most important simulations and computations is attached in Appendix C.

2 | Data description and models

The core business of Dow includes the production of plastic pellets. Plastic pellets are small grains of plastic that are used as raw material for producing end-user products. Packaging materials, computer components and parachute fabric are for example products that are made from plastic pellets such as polyethylene, urethane and ethylene-octene. Many of these production processes are highly efficient such that they are considered to be high-purity processes. In this chapter, we consider the data that originates from monitoring these high-purity processes for defects, where we take the production of plastic pellets as the leading example. First of all, we discuss the origin of the data in Section 2.1. Two distinct distributions are discussed afterwards, that are commonly used to model high-purity count data. It should be noted that due to a strict confidentiality policy at Dow, it has not been possible to include real plant data in this research. Instead, data properties have been thoroughly discussed with stakeholders inside of Dow, on which multiple simulation studies are based. Hence, data properties that are discussed throughout this report are based on simulations from models that capture the important features of the real data.

2.1 High-purity count data

Production of plastic pellets at Dow is a continuous process. At the end of the production line, production output is monitored to determine whether the produced pellets meet their design requirements. Pellets that contain any abnormalities or do not meet their standard are labelled as defects and fail the inspection. The process of producing and checking pellets for defects is executed as follows:

1. After production is complete, a representative of plastic pellets is collected and inspected continuously on minute basis.

2. The total amount of defects are counted by a detection algorithm, and reported in the monitoring data. For most production lines, defects are categorised based on their size. In such cases, multivariate monitoring data is obtained, reporting the amount of detected defects per size category are recorded each minute.

In this research project, we only consider univariate count data that represents the total amount of detected defects across all size categories. From now on, let us denote the total amount of detected defects with Y_i, at time $i = 1, 2, \ldots$. Then, Y_i is non-negative and integer such that we refer to it as count data. Producing plastic pellets at Dow is considered to be a high-purity process when the monitoring data Y_i includes mainly zero-defect occurrences.

However, some production lines of plastic pellets at Dow are more complex than described above. Namely, the production rate at which pellets are produced is not constant over time. This causes the amount of pellets in the detection stage to be inconstant over time as well. Hence, each sample of pellets that goes through the detection stage has different size. Therefore, the total weight of each batch is measured and reported. Let us denote the inspected weight with X_i for each point in time $i = 1, 2, \ldots$. It is trivial that the amount of detected defective pellets at time i depends on the inspected weight at

this particular time point. Hence, deviations in the inspected weight need to be taken into account when designing a monitoring system for the observed defect count.

In this thesis we consider monitoring methods for high-purity count data, where the response variable is affected by one input variable. In the context of plastic pellet production, we aim to monitor the total amount of detected defective pellets, while taking into account the inspected weight. Hence, we aim to detect contextual anomalies. Existing literature on this topic is discussed in Chapter 3. However, let us first consider two models for high-purity count data in the following section.

2.2 Models for zero-inflated count data

The Poisson and negative binomial distribution are common choices for modelling count data processes. The negative binomial distribution models the number of successes in a sequence of Bernoulli trials, before a specified number of failures. Let us denote number of failures with τ. The Poisson distribution is a limiting case of the negative binomial distribution, where $\tau = 1$ and the number of trials goes to infinity while the the expected value remains constant. This is stated by the Poisson limit theorem, that is explained in the textbook of Koralov and Sinai (2007) (Section 2.3). The negative binomial distribution is often considered as an alternative to the Poisson distribution, since it includes the size parameter τ. This size parameter allows for adjustable amounts of variation in the data, whereas the variance of the negative binomial distribution decreases when τ increases. Therefore, $1/\tau$ is sometimes referred to as the dispersion parameter.

It is explained in Section 2.1, that data from high-purity processes inherit a particularly large amount of zero observations. This is often referred to as zero-inflated (ZI) data. It is explained in Mahmood (2020), that modelling zero-inflated data with a Poisson distribution may cause violation of the equidispersion assumption, which leads to inaccurate estimations. The negative binomial distribution does also not account for an excess amount of zeros in the data, such that modelling zero-inflated data with a negative binomial distribution also leads to poor estimations due to overdispersion. Instead we can use zero-inflated distributions to model count data with an excessive amount of zeros. These distributions account for an additional proportion of zero occurrences with respect to the standard Poisson and negative binomial distribution. These models are therefore more appropriate to model count data for high-purity processes. The zero-inflated Poisson distribution and zero-inflated negative binomial distribution are discussed in the following sections.

2.2.1 The zero-inflated Poisson distribution

The zero-inflated Poisson (ZIP) distribution assumes that all observations emerge from two zero-generating processes. Namely, with probability $1 - p$, variable Y_i follows a Poisson distribution with expected value λ. With probability p we have that Y_i equals zero. This second process ensures the inflation of additional zeros to the Poisson model, such that we refer to these observations as structural zeros. The formal definition of the ZIP distribution for $i = 1, 2, \ldots$ is given by the following probability mass function.

$$P(Y_i = y) = \begin{cases} p + (1-p)e^{-\lambda} & \text{if } y = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y > 0 \end{cases} \tag{2.1}$$

Note that for $p = 0$, the ZIP distribution reduces to a regular Poisson distribution with parameter λ. The expected value of Y_i is defined as $E[Y_i] = (1-p)\lambda$, of which a proof is provided in A.1.1. Note that this is the $1 - p$ proportion of the expected value of a regular Poisson(λ) distribution, since the ZIP distribution only includes an additional point mass with value zero. In addition, the ZIP variance is defined as $\text{Var}(Y_i) = (1-p)(\lambda + p\lambda^2)$, of which a proof is provided in A.1.2. In the following chapters,

we will abbreviate the zero-inflated Poisson distribution with parameters p and λ to ZIP(p,λ). Here, p denotes the proportion of zero-inflation and λ is the expected value in case Y_i is not a structural zero.

2.2.2 The zero-inflated negative binomial distribution

The zero-inflated negative binomial (ZINB) distribution is constructed similarly to the ZIP distribution, and assumes that all observations emerge from two zero-generating processes. With probability $1-p$, the variable Y_i follows a negative binomial distribution with expected value λ and size parameter τ. With probability p we have that Y_i equals a structural zero. When using the Gamma notation of the negative binomial distribution, we can formally define the ZINB distribution for $i = 1, 2, \ldots$ in the following probability mass function.

$$P(Y_i = y) = \begin{cases} p + (1-p)\left(1 + \frac{\lambda}{\tau}\right)^{-\tau} & \text{if } y = 0 \\ (1-p)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}\left(1 + \frac{\lambda}{\tau}\right)^{-\tau}\left(1 + \frac{\tau}{\lambda}\right)^{-y} & \text{if } y > 0 \end{cases} \tag{2.2}$$

Size parameter τ is often chosen as an integer, but the ZINB distribution extends all non-negative real values with $\tau > 0$. Note that for $p = 0$, the ZINB distribution (2.2) reduces to the negative binomial distribution with parameter λ. The expected value of Y_i is defined as $E[Y] = (1-p)\lambda$, of which a proof is provided in A.1.3. In addition, the ZINB variance is defined as $\text{Var}(Y) = \lambda(1-p)(1+p\lambda+\lambda/\tau)$, of which a proof is provided in A.1.4. In the following chapters, we will abbreviate the zero-inflated negative binomial distribution to ZINB(p,λ,τ). Again, p denotes the proportion of zero-inflation, λ is the expected value of the negative binomial distribution in case Y_i is not a structural zero and τ is the size parameter.

2.3 Summary

The production of plastic pellets at Dow is considered to be a high-purity process. Monitoring these high-purity processes for defects results in univariate count data that that contains a large amount of zero observations. We denote the detected amount of defects with response variable Y_i, at time $i = 1, 2, \ldots$. The corresponding inspected weight is denoted with covariate X_i.

Since common count data distribution such as the Poisson and negative binomial distribution do not account for excessive amount of zeros, we should use a zero-inflated distribution to model the data. This can either be the zero-inflated Poisson distribution or the zero-inflated negative binomial distribution of which definitions are provided in Sections 2.2.1 and 2.2.2, respectively. Ultimately, we aim to construct a regression-based control chart for detecting contextual anomalies in monitoring data that originates from high-purity processes. Therefore, we focus on constructing regression-based control charts for data that follows a zero-inflated distribution. Literature on this topic is discussed in the following chapter.

3 Literature review

The field of statistical process control (SPC) was founded by Shewhart in 1924, when he published an internal report at the Bell Telephone Laboratories that contained the foundation of what is now called the Shewhart $\bar{X}$-chart. This control chart for normally distributed data is formally introduced in the paper of Shewhart (1925), but the ideas of Shewhart did not catch up with the American industry at the time. Instead, it was Deming and Juran that managed the breakthrough of SPC after Word War II (see e.g. Juran (1997) and Stauffer (2003)). Since then, many types of control charts have been introduced to fit specific detection goals or alternative distributions of the data. Some well-known examples are the cumulative sum chart from Page (1954), and the exponentially weighted moving average chart from Roberts (1959) and Shiryaev (1963). These charts are generally abbreviated to CUSUM and EWMA charts, respectively. Nowadays, SPC is rapidly gaining interest as our information infrastructures evolve and we aim to continuously analyse large amounts of data. The works of Megahed and Jones-Farmer (2015), Weese et al. (2016) and Qiu (2020) provide an interesting overview of monitoring an surveillance methods for big data applications that go beyond production line monitoring.

In this thesis, we focus on the application of regression-based control charts for high-purity processes. The first regression-based control chart was introduced by Mandel (1969), for the purpose of detecting contextual abnormal behaviour in normally distributed data. This topic appears in econometrics as monitoring structural change, see e.g. Chu et al. (1996). The studies of Brown et al. (1975) and Dufour (1982) have introduced a recursive approach for regression-based control charts. Both studies are discussed in Section 3.1, where a literature review is presented regarding the established monitoring techniques for detecting contextual anomalies. Regression models for zero-inflated count data are discussed in Section 3.2, after which a literature overview regarding GLM-based control charts is provided in Section 3.2. Finally, a solution strategy for this thesis is presented in Section 3.4.

3.1 Monitoring techniques

Originally, regression-based control charts were introduced to monitor the deviance from an established regression model over time. In this case, the regression model is estimated from a stable period in the process, referred as the Phase I period by Hawkins et al. (2003). Monitoring residuals that are obtained from a fixed Phase I regression model is also referred to as monitoring predictive residuals by Van Dalen (2018). Here, it is shown that for normally distributed observations, that predictive residuals are correlated and therefore dependent, since all residuals are obtained from the same regression model. The application of predictive residuals in a control chart is therefore violating the assumption of independent observations.

The research from Brown et al. (1975) introduces a recursive approach for obtaining regression residuals over time. Here, a new linear regression model is fitted at the arrival of each new observation, after which the regression residuals are obtained. These residuals are referred to as recursive residuals, and are proved to be uncorrelated with zero mean and constant variance for normally distributed observations. Monitoring recursive residuals dismisses the need for a stable Phase I period, since the recursive approach

can start after collection of the first $p+1$ data points. In this context, p denotes the number of covariates in the regression model (see Van Dalen (2018)).

In addition to recursive residuals, Dufour (1982) introduces an approach for obtaining recursive regression coefficients. A new linear regression model is fit to the data upon the arrival of each new observation, after which the estimated regression coefficients are obtained. Dufour (1982) proves that the difference between consecutive regression coefficients follows an independent and identical normal distribution. Chu et al. (1996) and Zeileis et al. (2005) use this property to introduce a monitoring scheme where the differences between recursively estimated regression coefficients are applied in a control chart.

All together, multiple monitoring techniques have been introduced for regression-based control charts. However, all previously mentioned studies consider normally distributed data for which linear regression models are applied. In this thesis, we are dealing with zero-inflated count data such that linear regression models do not apply. The following section provides an overview of regression models for zero-inflated count data.

3.2 Regression models for zero-inflated count data

Generalized linear models (GLM) are a special class of regression models that originate from Nelder and Wedderburn (1972). These regression models go beyond the normal distribution and extend to all distributions in the exponential family. Jørgensen (1997) expands this class of distributions, and states that GLMs apply when the response variable follows a distribution that belongs to the family of exponential dispersion models (EDM). This family includes both continuous and discrete distributions such as the normal, Gamma and Poisson distribution. However, zero-inflated distributions do not belong to the EDM family such that GLM regression models are not defined for these distributions. To overcome this, Lambert (1992) introduced a custom regression model for the zero-inflated Poisson (ZIP) distributed data. Here, it is assumed that each observation Y_i follows a ZIP(p_i,λ_i) distribution, where parameter p_i and λ_i are functions of the model covariates. Lambert (1992) denotes the regression models where parameter p_i and λ_i are affected by distinct sets of covariates as the ZIP model. In case p_i and λ_i depend on the same set of covariates, then the regression model is denoted with ZIP(τ). The additional notation with τ indicates that every p_i can be written as a function of λ_i. Parameter estimations of both models are proved to be asymptotically normal and extensive simulations show that estimates can be trusted when the ZIP or ZIP(τ) are fitted on sufficiently large data sets, i.e. $n \geq 100$. For parameter estimation of the ZIP model, convergence of the expectation maximisation (EM) algorithm is proved.

Heilbron (1994) introduced a zero-inflated negative binomial (ZINB) regression model, as a generalisation to the ZIP model. In this study, parameter estimates are again proved to have asymptotically normal distribution. Application of the ZINB model in a use case with zero-inflated data illustrates the improved model fit in terms of increased log-likelihood, with respect to standard negative binomial regression. Both models from Lambert (1992) and Heilbron (1994) are similar to GLM, except for the fact that different algorithms are applied for parameter estimation. For GLMs, the iterative weighted least squares (IWLS) algorithm is applied, whereas the EM algorithm is used for parameter estimation of the ZIP model. The Broyden–Fletcher–Goldfarb –Shanno (BFGS) algorithm is applied for the ZIP(τ) and ZINB regression model. A more detailed discussion on the ZIP and ZINB regression models and their similarity with GLM is provided in Chapter 6.

Over the years, many studies have been conducted on monitoring residuals from linear regression models. However, GLM-based control charts are less well researched. Literature regarding regression-based control charts with the zero-inflated (ZI) regression models is even more scarce, and only appears in very recent studies. Hence, we consider literature regarding both GLM-based control charts and ZI regression-based control charts in this thesis. An overview of literature regarding GLM- and ZI regression-based control charts is provided in the next section.

3.3 GLM- and ZI regression-based control charts

The first GLM-based control charts were introduced around the same time by Skinner et al. (2003) and Jearkpaporn et al. (2003). Skinner et al. (2003) introduces a GLM-based Shewhart chart with deviance residuals, for Poisson distributed response variables. Control limits of the GLM-based Shewhart chart are solved to match the in-control (IC) false alarm rate of the c-chart, in order to ensure fair performance comparison. A simulation study shows that the GLM-based control chart outperforms the traditional c-chart for both univariate and multivariate Poisson distributed data, in terms of lower out-of-control average run length, i.e., ARL_1. As an extension to the research in 2003, Skinner et al. (2004) compares the performance of linear ordinary least squares (OLS) regression-based Shewhart charts with GLM-based Shewhart charts. The data is assumed to follow an overdispersed Poisson distribution that is modelled by a Poisson mixture model with additional dispersion parameter. Again, control limits of the OLS and GLM-based control charts are solved to match the IC false alarm rate of the c-chart. Simulations show that GLM-based Shewhart chart outperform both OLS-based Shewhart chart and standard Shewhart c-chart, in terms of ARL_1, when aiming to detect contextual anomalies. The fact that the GLM-based Shewhart chart outperforms the OLS-based Shewhart chart proves that ignoring non-normality of the data when constructing a regression-based control chart, leads to less accurate performance in terms of average run length.

Jearkpaporn et al. (2003) introduces a GLM-based Shewhart chart with deviance residuals, for Gamma distributed data. It is shown that in the particular case of Gamma GLM, the deviance residual is defined as a likelihood ratio statistic which is approximately normal. Performance Gamma GLM-based Shewhart chart is compared with traditional Shewhart chart for individual observations. Control limits of both charts are solved to obtain a specified in-control ARL, i.e., ARL_0. Simulations show that the Gamma GLM-based Shewhart chart outperforms the traditional Shewhart chart in terms of lower ARL_1. The results in Skinner et al. (2003), Skinner et al. (2004) and Jearkpaporn et al. (2003) combined supports the general perception that, regression-based control charts outperform traditional control charts in terms of ARL_1, when aiming to detect contextual anomalies.

In addition to the charts for Poisson and Gamma distributed data, Park et al. (2018) introduced regression-based Shewhart charts for negative binomial and Conway-Maxwell (COM) Poisson distributions. Interesting from this study is the addition of principal-component-analysis (PCA) as a method to transform a large number of possibly correlated covariates into a smaller number of uncorrelated covariates. Also, control limits are fixed at 3σ from the centre line for all regression-based Shewhart charts, and ARL_1 performance is evaluated for underdispersed, equidispersed and overdispersed data, generated by the COM-Poisson distribution with various values for dispersion parameter $v = 0.8, 1, 1.5$ respectively. In case of underdispersed data, COM-Poisson and Poisson regression-based control charts outperform the negative binomial regression-based control chart in terms of lower ARL_1. Similarly, the negative binomial an COM-Poisson regression-based control charts outperform the Poisson GLM-based control chart. Hence, it can be concluded from Park et al. (2018) that Poisson GLM-based control charts perform better for underdispersed count data, whereas the negative binomial regression-based control chart perform better in case of overdispersed data.

Mahmood (2020) focuses on GLM-based control charts for high-purity processes. A regression-based Shewhart chart for zero-inflated Poisson (ZIP) distributed data is introduced, as well as a regression-based control chart for data that follows a zero-inflated negative binomial (ZINB) distribution. The regression models of Lambert (1992) and Heilbron (1994) are applied to the charts, after which Pearson residuals are obtained. The use of Pearson residuals is remarkable in the research, since deviance residuals are commonly used in literature.

In general it is often assumed that residuals from a well-fitted regression model are normally distributed. It is however stated in Dunn and Smyth (2018) (Section 8.3), that when the response data is dependent and non-normal, as is the case with GLM regression, normality of Pearson and deviance

residuals is unlikely. In particular when modelling discrete outcome variables when the variance of the response is high. Quantile residuals are introduced in Dunn and Smyth (1996) as an alternative to Pearson and deviance residuals. By definition of the probability integral transform, their distribution is exactly normal apart from the sampling variability in estimating distribution parameters. This implies that quantile residuals are exactly normally distributed in case all parameters are estimated at their true values. A deviation from normality can be observed for less well-fitted models. For this reason it is recommended in Dunn and Smyth (2018) (Section 8.3.4), to use quantile residuals when accessing the goodness of fit for regression models with discrete data. An additional advantage of quantile residuals is that their application in a Shewhart chart is not violating the normality assumption. The question remains, which type of residuals should we apply when constructing a regression-based control chart?

Recent studies from Park et al. (2020) and Jamal et al. (2021) address this matter by comparing control chart performances for various GLM-based control charts with different residual types. Park et al. (2020) introduces regression-based Shewhart charts with quantile residuals, in case the data follows a normal, Poisson, binomial, negative binomial, Conway-Maxwell-Poisson and zero-inflated-Poisson distribution. The performance of these charts in compared with similar regression-based control charts with deviance residuals. The control limits of the Shewhart charts are set at $w \cdot \sigma$ where $w = 1, 2, 3$. ARL_1 simulations show for all distributions except the Binomial distribution, that GLM-based Shewhart charts with quantile residuals outperform similar charts with deviance residuals. These results of Park et al. (2020) are however contradicted by the research of Jamal et al. (2021). This study evaluates the performance of GLM-based Shewhart, EWMA and CUSUM charts for Conway-Maxwell-Poisson distributed data in terms of simulated ARL_1 performance. Charting constants in the Shewhart, EWMA, and CUSUM chart were solved to obtain pre-specified ARL_0 values to ensure fair comparison between the charts with different residual types. The results indicate that the control charts with deviance residuals outperform the charts with quantile residuals.

The contradicting conclusions in recent studies of Park et al. (2020) and Jamal et al. (2021) debate the existence of one single monitoring strategy that is superior to all others. It rather suggests that the optimal performing residual type in a regression-based control chart should be found according to specific properties of the monitoring data and the detection goal. Additionally we can state to the best of our knowledge, that there is not yet any research published regarding performance of all three residual types in regression-based control charts for high-purity processes. Mahmood (2020) provides performance results of the ZIP and ZINB Shewhart chart with Pearson residuals. Park et al. (2020) provides ARL_1 results of the ZIP Shewhart chart with deviance and quantile residuals. However, results from both studies cannot be compared, since control limits are obtained differently. Hence, we can conclude from the literature review that there exists a research gap regarding the performance of Pearson, deviance and quantile residuals in regression-based control charts for zero-inflated data.

As a final note, it is remarkable that results from Mahmood (2020) and Park et al. (2020) are obtained under the assumption ZIP and ZINB regression coefficients are known. The provided results illustrate the baseline performance of the ZIP and ZINB Shewhart chart with Pearson residuals, while ignoring the effects of Phase I estimation. It is known that the poor Phase I estimation can cause the true ARL_0 to be much lower than the intended ARL_0, in case of a Shewhart chart with normally distributed data, see e.g. Albers and Kallenberg (2004). It is shown in Shu et al. (2005) that such effects of Phase I estimation are also true when constructing linear regression-based control charts for normally distributed data. Therefore, it is also expected that Phase I estimation affect the ARL_0 performance of the regression-based control charts for zero-inflated data, although this remains unquestioned in Mahmood (2020) and Park et al. (2020). This thesis aims to provide more insight in the performance of Pearson, deviance and quantile residuals in a regression-based control chart for zero-inflated data. In addition, we also aim to estimate the effects of Phase I estimation on the performance of regression-based control charts for zero-inflated data. The solution strategy is provided in the next section.

3.4 Solution strategy

In this thesis, we will focus on monitoring predictive Pearson, deviance and quantile residuals in regression-based control charts for high-purity processes. There is no literature available on monitoring recursive residuals or recursive regression coefficients for high-purity processes, such that recursive monitoring is out of scope in this project as well. The goal is to provide Dow with a framework for identifying the best performing residual type in a regression-based control chart for zero-inflated data. More specifically, we narrow the scope of this research to ZIP and ZINB regression-based Shewhart charts.

To become acquainted with monitoring zero-inflated data, we start with the design of a simplified monitoring scheme where the response variable is not affected by any covariates. The ZIP-EWMA control chart for independent and identically distributed observations is proposed for this purpose in Chapter 5. Here, it is explored how the ZIP-EWMA performance is affected by various proportions of zero-inflated. Afterwards, we proceed with the more advanced setting, where the response variable is affected by one covariate. The ZIP and ZINB regression-based Shewhart chart are introduced in Chapter 7 with predictive Pearson, deviance and quantile residuals. Each regression-based control chart is evaluated under various proportions of zero-inflated in the IC Phase I data, to identify which chart performs best for each scenario.

Monitoring predictive residuals requires a Phase I to obtain estimates of distributional parameters. Collecting a Phase I data set that reflects a stable period does not represent a limitation at Dow, since historical monitoring data is available in abundance. However, it is mentioned in Section 3.2 that poor Phase I estimation can cause the true ARL_0 of a control chart to be much lower than the intended ARL_0. In order evaluate the effect of Phase I estimation for the ZIP and ZINB regression-based Shewhart chart, we propose two distinct performance analysis strategies. At first, the baseline performance of the ZIP and ZINB regression-based Shewhart charts is established under the assumption that all regression parameters are known. Afterwards, we evaluate the performance of the same charts when using regression models that are estimated from simulated Phase I data. The methodology for both approaches are described in detail in Chapter 7. A graphical representation of the solution strategy is shown in Figure 3.1, where the dotted lines represent the objective of this thesis.

3.5 Summary

It is discussed in Section 3.1 that since Mandel (1969), various monitoring techniques have been introduced for detecting contextual anomalies in monitoring data. These techniques include monitoring predictive residuals, recursive residuals and recursive regression coefficients. The ZIP and ZINB regression models from Lambert (1992) and Heilbron (1994) are discussed in Section 3.2 and GLM-based control charts are discussed afterwards in Section 3.3. The literature review shows that regression-based control charts outperform traditional control charts, when aiming to detect contextual anomalies and that GLM-based control charts outperform linear regression-based control charts, when dealing with non-normal data. Regression-based Shewhart charts for zero-inflated data are introduced by Mahmood (2020), for ZIP and ZINB distributed data. The use of Pearson residuals in this paper is exceptional, since deviance residuals are commonly used in literature. Additionally, Park et al. (2020) and Jamal et al. (2021) illustrate the possibility of using quantile residuals for monitoring as well. However, a clear consensus remains absent regarding which residual type performs best in monitoring schemes for high-purity count data.

The solution strategy of this thesis is provided in Section 3.4, of which a graphical representation is shown in Figure 3.1. It is proposed to start with performance evaluation of the ZIP-EWMA chart, for monitoring high-purity count data without covariates. Then, the ZIP and ZINB regression-based Shewhart charts are evaluated for monitoring high-purity count data with covariates. Pearson, deviance and quantile residuals are applied, and performance is tested for various proportions of zero-inflation in the data. Definitions for the Shewhart and EWMA chart are provided in the following chapter.

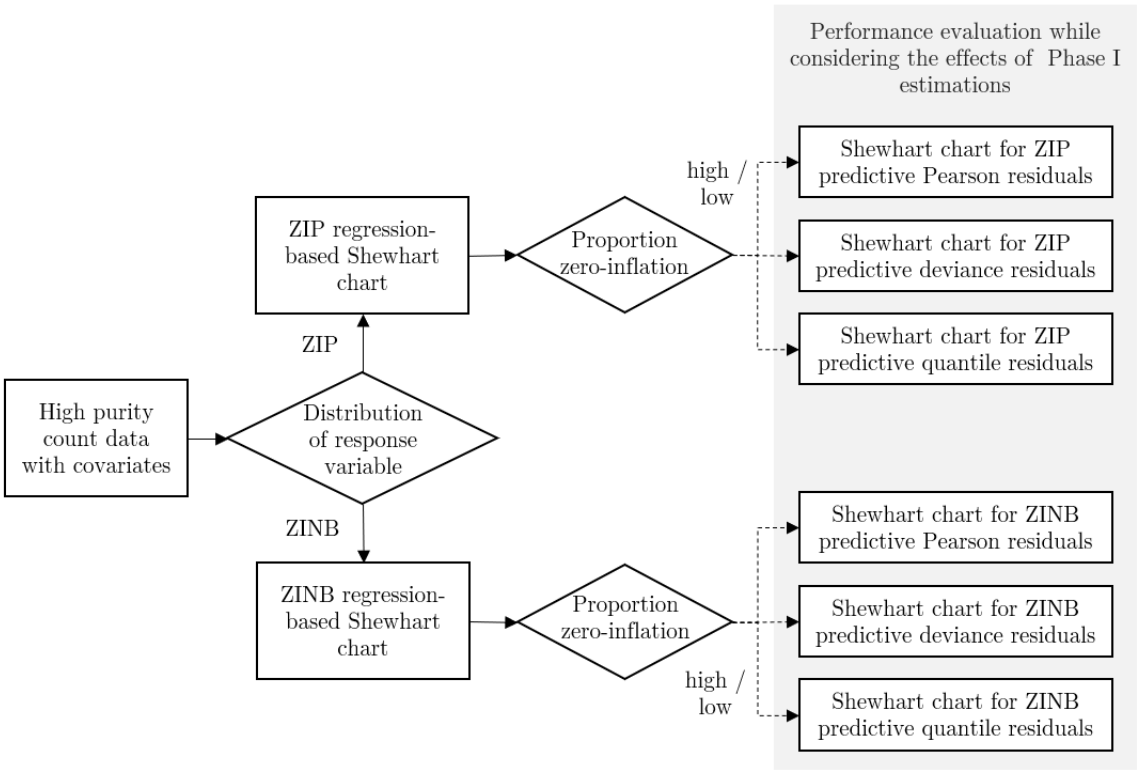

Figure 3.1: Flow chart of the solution strategy.

4 | Statistical process control

Statistical process control (SPC) is the field that employs statistical tools and methods to control the quality of industrial processes. Statistical tools that are considered in SPC include design of experiment, process capability analysis, regression and many more. In particular, SPC control charts are highly effective for detecting distributional shifts in process performance data. In this thesis, we focus on monitoring methods for high-purity count data.

It is described in the solution strategy of Section 3.4, that we aim to construct both a ZIP-EWMA chart, as well as ZIP and ZINB regression-based Shewhart charts. Therefore, we elaborate on SPC methodologies in this chapter. The main concepts in SPC are discussed in Section 4.1, followed by a definition of the Shewhart and EWMA charts in Section 4.2. The textbook of Qiu (2013) is used as a reference throughout this chapter.

4.1 Concepts in statistical process control

Shewhart (1925) introduced the control chart to distinguish between common cause and special cause variation. Common cause variation is the type of variability that is caused by uncontrollable factors, which are inherited in the process. It is explained in Qiu (2013) (Section 1.3), a process is considered to be in control (IC) when all variability is due to common cause. Special cause variation is the variability due to unexpected circumstances, which are not inherited in the process. A process is considered to be out-of-control (OC) when process output contains special cause variation. Control charts aim to identify special cause variation, as soon as it occurs in the process. Early detection of process malfunctions contributes to more efficient manufacturing, which typically leads to lower overall production costs and more efficient use of materials.

SPC is usually divided into two phases, see e.g. Qiu (2013) (Section 1.3). In Phase I, the process is explored and adjusted to make it run stably. The goal is to obtain a stable data set, which represents the IC process. Involved parameters are estimated from the Phase I data, which benchmark a stable process in the future. In Phase II it is assumed that the process runs stably from the beginning. Hence, process data is assumed to follow the IC distribution that is established in Phase I. The goal of Phase II is to monitor the process continuously over time, to make sure that it keeps running stably. SPC control charts are statistical tools that achieve the Phase II goal. Hence, they monitor process data over time, to identify when it becomes OC. Definitions of the univariate Shewhart and EWMA chart are presented in the following section.

4.2 SPC Control charts

Control charts are statistical tools for repeated hypothesis testing. At the arrival of each new observation, we test whether the process is still in control. The null hypothesis of each repetition states that the process is IC and the alternative hypothesis states that the process is OC. Let us consider an industrial process, from which a quality measurement is obtained at each point in time, e.g. each minute. Let us define the

consecutive quality observations as $\{Y_1, Y_2, \dots\}$. We assume that all observations follow an independent and identical normal distribution with mean μ_0 and standard deviation σ_0, in case the process is in control. At the arrival of each observation Y_i with $i = 1, 2, \dots$, a control chart tests whether the process is IC. Hence, the control chart hypotheses in this example are defined as

$$H_0: \quad Y_i \sim N(\mu_0, \sigma_0^2) \quad \text{for} \quad i = 1, 2, \dots$$

$$H_1: \begin{cases} Y_i \sim N(\mu_0, \sigma_0^2) & \text{for} \quad i = 1, \dots, \mathcal{T} \\ Y_i \sim N(\mu_1, \sigma_1^2) & \text{for} \quad i = \mathcal{T}+1, \mathcal{T}+2, \dots. \end{cases}$$

where $\mu_1 \neq \mu_0$ and or $\sigma_1 \neq \sigma_0$. In this notation, time $\mathcal{T}$ is referred to as the changepoint by Hawkins et al. (2003), after which the process becomes OC. The statistical tests are executed by plotting the charting statistic, an upper control limit (UCL) and a lower control limit (LCL) over time. As long as the charting statistic stays in between the control limits, the null hypothesis is not rejected. Hence, it is concluded that the process is IC and the monitoring continues. When the charting statistic exceeds either one of the control limits, the null hypothesis is rejected and it is concluded that the process is OC. In this case, an OC signal is produced and the production process is stopped for evaluation. Figure 4.1 shows an example of a Shewhart chart.

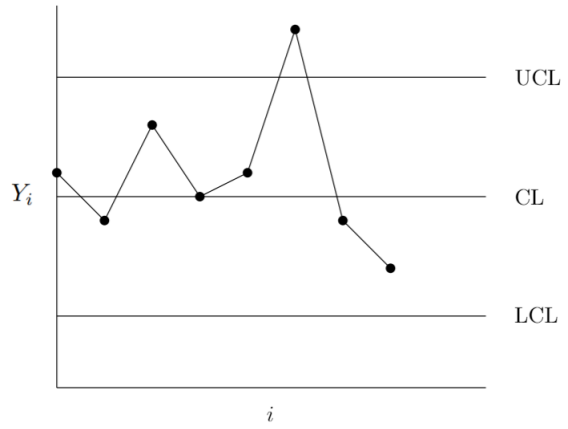

Figure 4.1: Example of a Shewhart control chart.

It is possible that the alternative hypothesis is not rejected while the process is actually IC. This is referred to as a false alarm. It is common to evaluate the IC performance of a control chart by its run length (RL). The RL is the total number of observations from the initial time point until the first OC signal. The RL is a random variable since it depends on random observations. The average run length (ARL) is therefore often used as a performance measure for SPC control charts in literature. The ARL of an IC process is denoted with ARL_0, and a high ARL_0 is desired. However, the run length distribution can be highly skewed such that the standard deviation of the run length (SDRL) is often also considered as a performance measure. The SDRL of an IC process is denoted with $SDRL_0$ and a low $SDRL_0$ is desired.

Now let us consider a shift in the distribution of observations, such that process that becomes OC. The number of observations from the time of the shift occurrence to the time of signal is referred to as the OC run length. The average OC run length is denoted with ARL_1 and the standard deviation of the OC run length is denoted with $SDRL_1$. It is ideal to have a fast OC signal when the process becomes OC, such that a low ARL_1 and a low $SDRL_1$ are desired. It is common in SPC literature to construct control limits in order to achieve a satisfactory ARL_0, and evaluate the charts performance based upon its ARL_1. Construction of control charts is discussed in the following sections. A definition

of the Shewhart chart is provided in Section 4.2.1, and the EWMA chart is defined in Section 4.2.2.

4.2.1 The Shewhart chart

The first SPC control chart is introduced in Shewhart (1925), and is therefore referred to as the Shewhart chart. Let us again define observations as $\{Y_1, Y_2, \dots\}$, that are independent and identical normally distributed with mean μ_0 and standard deviation σ_0 in case the process is IC. The control limits are fixed at equal distance from the centre line (CL). This distance is determined by charting constant L, and the IC standard deviation σ_0. Hence, the Shewhart chart is defined as

$$\begin{aligned} UCL &= \mu_0 + L \cdot \sigma_0 \\ CL &= \mu_0 \\ LCL &= \mu_0 - L \cdot \sigma_0. \end{aligned} \tag{4.1}$$

Values for μ_0 and σ_0 can be estimated from the Phase I data in case the true parameters are unknown. Under the assumption of independent and identical normally distributed observations, it is usually chosen to fix $L = 3$, which leads to a false alarm probability $\alpha = 0.0027$. The IC run length follows a geometric distribution in this case, with probability parameter α. Therefore, we obtain $ARL_0 = 1/0.0027 \approx 370$. In addition, the $SDRL_0$ equals $\sqrt{1-\alpha}/\alpha \approx 370$ in this case. It is described in Qiu (2013) (Section 3.2) that for small values of α we have $ARL_0 \approx SDRL_0$.

Besides $L = 3$, other values for L can also be chosen when aiming to achieve a different ARL_0. The chart as defined in (4.1) is sometimes referred to as a Shewhart chart with symmetric control limits, since the UCL and LCL are symmetric around the centre line. Independent and identically distributed observations $\{Y_1, \dots, Y_n\}$ have equal probability of exceeding the UCL and LCL in (4.1), as long as their distribution is symmetric. However, this assumption is often violated in practice. If the true distribution of the data is skewed, then the control limit towards which the data is skewed is less likely to be exceeded.

According to Xie et al. (2002a) (Section 2.1), probability control limits can be used instead in such cases. Let us assume observations $\{Y_1, \dots, Y_n\}$ are i.i.d. distributed with cumulative distribution function F_θ. Here, θ is a defined set of parameters. Then, for a certain quantile level α, we can determine the upper and lower control limit as $Q_1 : F_\theta(Q_1) = 1 - \alpha/2$ and $Q_2 : F_\theta(Q_2) = \alpha/2$ respectively. The Shewhart control chart with probability limits is then defined as

$$\begin{aligned} UCL &= Q_1 \\ CL &= \mu_0 \\ LCL &= Q_2. \end{aligned} \tag{4.2}$$

The run length distribution is again geometrically distributed with probability parameter α, as described by Chakraborti (2007). Hence, for $\alpha = 0.0027$ it holds again that $ARL_0 \approx SDRL_0 \approx 370$. In case distributional parameters θ are unknown, Q_1 and Q_2 can be solved based upon the Phase I estimation of F_θ. Figure 4.2 illustrates the way in which the symmetric Shewhart chart as defined in (4.1) are a better fit for symmetrically distributed data, while the Shewhart chart with quantile limits is better suited for observations with skewed distribution.

The fundamental difference between a Shewhart chart and an EWMA or CUSUM is that the Shewhart chart includes current observations only to test for special cause variation. EWMA and CUSUM charts both include all observations for testing, from the current time point and earlier time points. In addition, EWMA charts assign a weight allocation with exponential decay to the historical observations, while CUSUM charts include all historical data with equal weight. For normally distributed variables Y_i we know that Shewhart charts are better in detecting large distributional shifts in process performance, whereas EWMA and CUSUM charts are better in detecting small and persistent shifts. It should be

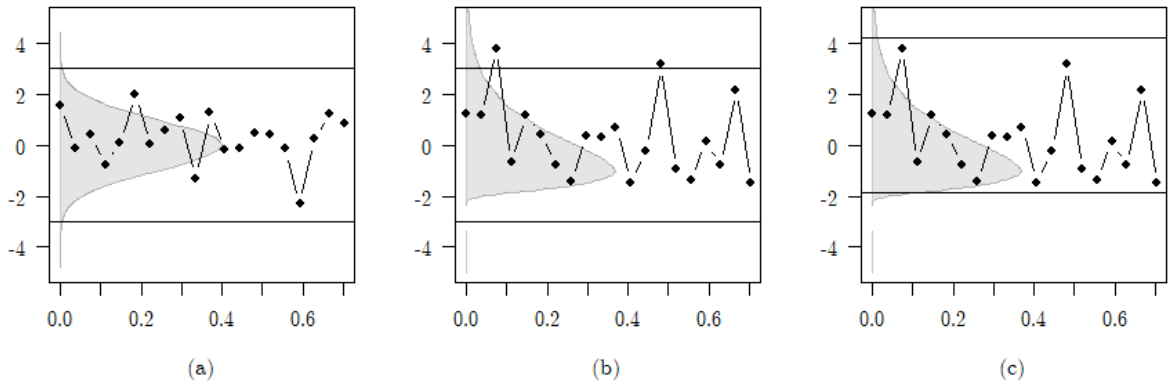

Figure 4.2: Graphical representation of: (a) a symmetric Shewhart chart with normally distributed charting statistic, (b) a symmetric Shewhart chart with skewed charting statistic, (c) a Shewhart chart with probability limits and skewed charting statistic.

noted that this is not necessarily true when Y_i follows a non-normal distribution, but Shewhart chart are nevertheless often employed for detecting large distributional shifts in the data. A definition of the EWMA chart is provided in the following section.

4.2.2 The EWMA chart

Let us again define observations as $\{Y_1, Y_2, \dots\}$, that are independent and identically distributed with mean μ_0 and standard deviation σ_0 in case the process is IC. The charting statistic of the EWMA chart is the exponentially weighted moving average of all historical observations up to the current time point. Let us define the current time point as time n, such that observations $\{Y_1, \dots, Y_n\}$ are collected so far. When using notation from Qiu (2013) (Section 5.2), charting statistic E_n at time n is defined as

$$E_n = wY_n + (1-w)E_{n-1} \tag{4.3}$$

for $n = 1, 2, \dots$. Here, $E_0 = \mathrm{E}[Y_0] = \mu_0$ and w is the weight parameter with $0 < w \leq 1$. The contribution of past observations to E_n is large when w is chosen close to zero. On the other hand, more weight is assigned to current observations for larger values of w. The EWMA chart reduces to a Shewhart chart for $w = 1$. The EWMA control limits at time n are defined as

$$\begin{aligned} UCL_n &= \mathrm{E}[E_n] + L\sqrt{\mathrm{Var}(E_n)} \\ CL &= \mathrm{E}[E_n] \\ LCL_n &= \mathrm{E}[E_n] - L\sqrt{\mathrm{Var}(E_n)}. \end{aligned} \tag{4.4}$$

for $n = 1, 2 \dots$. By iterating (4.3) we can write

$$E_n = (1-w)^n \mu_0 + w \sum_{i=0}^{n-1} (1-w)^i Y_{n-i}. \tag{4.5}$$

It follows that E_n is a convex combination of all observations Y_i for $i = 1, \dots, n$ since by definition of the geometric sum we have

$$(1-w)^n + w \sum_{i=0}^{n-1} (1-w)^i = (1-w)^n + w \left(\frac{1-(1-w)^n}{w} \right) = 1.$$

By using definition (4.5), the expected value of E_n is defined as

$$\begin{aligned} \mathrm{E}\left[E_n\right] &= \mathrm{E}\left[(1-w)^n\mu_0 + w\sum_{i=0}^{n-1}(1-w)^i Y_{n-i}\right] \\ &= (1-w)^n\mu_0 + w\sum_{i=0}^{n-1}(1-w)^i\mu_0 = \mu_0. \end{aligned} \tag{4.6}$$

Note that $\mathrm{E}\left[E_n\right] = \mu_0$ for all $n = 1, 2, \ldots$ since E_n is a convex combination of all i.i.d. observations Y_i with $i = 1, \ldots, n$. In addition to the expected value, the variance of E_n is defined as

$$\mathrm{Var}(E_n) = \mathrm{Var}\left((1-w)^n\mu_0 + w\sum_{i=0}^{n-1}(1-w)^i Y_{n-i}\right) = w^2\sum_{i=0}^{n-1}(1-w)^{2i}\mathrm{Var}(Y_{n-i}) = \frac{w\left(1-(1-w)^{2n}\right)\sigma_0^2}{2-w}$$

where definition (4.5) is substituted for E_n in the first step, and it is used in the second step that all observations Y_i with $i = 1, \ldots, n$ are independent from each other. With these definitions for $E[E_n]$ and $\mathrm{Var}(E_n)$, the control limits of the EWMA chart at time n are defined as

$$\begin{aligned} UCL_n &= \mu_0 + L\sqrt{\frac{w}{2-w}\left(1-(1-w)^{2n}\right)\sigma_0^2} \\ CL &= \mu_0 \\ LCL_n &= \mu_0 - L\sqrt{\frac{w}{2-w}\left(1-(1-w)^{2n}\right)\sigma_0^2}. \end{aligned} \tag{4.7}$$

The limits of the EWMA chart converge for large values of n. Hence, the asymptotic control limits for $n \to \infty$ are given by

$$\begin{aligned} UCL_\infty &= \mu_0 + L\sqrt{\frac{w}{2-w}}\sigma_0 \\ CL &= \mu_0 \\ LCL_\infty &= \mu_0 - L\sqrt{\frac{w}{2-w}}\sigma_0. \end{aligned} \tag{4.8}$$

Charting constant L is obtained to achieve a predefined ARL_0 value. For Shewhart charts it is described that the RL distribution is geometric, since the charting statistic is independent of time. However, the EWMA charting statistic includes all historical observations, such that E_{n-1} and E_n are not independent for $n = 2, 3, \ldots$. Hence, a geometric run length distribution does not apply here and a closed form definition ARL does not exist. However, values of ARL_0 and $SDRL_0$ can be obtained by simulation, for specified values of w, L and a known distribution of observations $\{Y_1, Y_2, \ldots\}$. Knoth (2021) provides the `spc` package in `R` for simulating the EWMA ARL_0, for various distributions of observations.

4.3 Summary

In this chapter, we discussed the basic concepts in the field of statistical process control in Section 4.1. The hypotheses and general structure of a SPC control chart are discussed in Section 4.2, where the ARL and SDRL performance measures are introduced as well. A definition of the Shewhart chart is provided in Section 4.2.1, where both symmetric and probability control limits are discussed. Finally, a definition for the EWMA chart is provided in Section 4.2.2. In the following chapter, we will continue with an application of the EWMA chart for ZIP distributed data that is not affected by any covariates. The application of ZIP and ZINB regression-based Shewhart charts for data that is affected by a covariate is discussed afterwards in Chapter 7.

5 Monitoring count data without covariates

The goal of this thesis is to design control charts for detecting contextual anomalies in zero-inflated univariate count data. However, not all high-purity processes at Dow have a covariate that affects the variable of interest. Therefore, we start with the design of a monitoring scheme where the data is not affected by any covariates. An EWMA control chart for independent and identically ZIP distributed observations is proposed in this chapter. Here, it is explored how the ZIP-EWMA performance is affected by various proportions of zero-inflated in the in-control monitoring data.

The hypotheses and definition of a ZIP-EWMA chart are provided in Section 5.1. Then, the computation of control limits and performance evaluation is discussed in Section 5.2. Finally, discussion of this intermediate ZIP-EWMA study are discussed in Section 5.4, followed by the conclusion. Throughout this chapter, the textbook of Qiu (2013) is used as a main reference. All computations are conducted in `R`, and the code is attached in Appendix C.1.

5.1 The ZIP-EWMA control chart

In this section we define a two-sided exponentially weighted moving average control chart for ZIP distributed observations. In the context of producing pellets, we denote the total amount of detected defects with $\{Y_1, Y_2, \dots\}$ for each point in time $i = 1, 2, \dots$. Let us assume that all observations follow an independent and identical ZIP distribution with parameters p_0 and λ_0, in case the process is IC. The ZIP distribution is provided in Section 2.2.1, where it is also defined that $E[Y_i] = (1-p_0)\lambda_0$ and $\text{Var}(Y_i) = (1-p_0)(\lambda_0 + p_0\lambda_0^2)$ when the process is stable and in control. When using the same notation as in Section 4.2, the hypotheses of a ZIP control chart can be formally defined as

$$H_0: \quad Y_i \sim ZIP(p_0, \lambda_0) \quad \text{for} \quad i = 1, 2, \dots$$

$$H_1: \begin{cases} Y_i \sim ZIP(p_0, \lambda_0) & \text{for} \quad i = 1, \dots, \mathcal{T} \\ Y_i \sim ZIP(p_1, \lambda_1) & \text{for} \quad i = \mathcal{T}+1, \mathcal{T}+2, \dots. \end{cases}$$

where $p_1 \neq p_0$ and or $\lambda_1 \neq \lambda_0$, and $\mathcal{T}$ is the changepoint after which the process becomes out-of-control. Definitions of the EWMA charting statistic and control limits are provided in Section 4.2.2. The ZIP-EWMA charting statistic is defined as in (4.3), and the control limits are constructed by substituting (5.1) with $\mu_0 = E[Y_i] = (1-p_0)\lambda_0$ and $\sigma_0^2 = \text{Var}(Y_i) = (1-p_0)(\lambda_0 + p_0\lambda_0^2)$. When denoting the current

time point as time n, the ZIP-EWMA chart is defined by

$$\begin{aligned} UCL_n &= (1-p_0)\lambda_0 + L\sqrt{\frac{w}{2-w}\left(1-(1-w)^{2n}\right)(1-p_0)(\lambda_0+p_0\lambda_0^2)} \\ CL &= (1-p_0)\lambda_0 \\ LCL_n &= \max\left\{0, (1-p_0)\lambda_0 - L\sqrt{\frac{w}{2-w}\left(1-(1-w)^{2n}\right)(1-p_0)(\lambda_0+p_0\lambda_0^2)}\right\} \end{aligned} \tag{5.1}$$

The lower control limit is defined as the maximum between 0 and the unrestricted lower control limit value since the ZIP distribution has a positive solution space. The asymptotic control limits for $n \to \infty$ are given by

$$\begin{aligned} UCL &= (1-p_0)\lambda_0 + L\sqrt{\frac{w}{2-w}(1-p_0)(\lambda_0+p_0\lambda_0^2)} \\ CL &= (1-p_0)\lambda_0 \\ LCL &= \max\left\{0, (1-p_0)\lambda_0 - L\sqrt{\frac{w}{2-w}(1-p_0)(\lambda_0+p_0\lambda_0^2)}\right\} \end{aligned}$$

The charting constant L is unknown, and can be obtained by solving the control limit equation in (5.1) to achieve a predefined ARL_0 value.

5.2 Constructing the ZIP-EWMA chart

In the SPC literature, it is common to construct a control chart for a fixed ARL_0 value, and evaluate the corresponding ARL_1. This strategy is also followed to construct the ZIP-EWMA chart. The run length is a random variable such that we can define the ARL as the expected value of RL. However, it is described in Section 4.2.2 that a closed form definition of the EWMA ARL does not exist. Monte Carlo simulations can be used to numerically obtain the performance metrics, as explained in Qiu (2013) (Section 5.2). It is assumed throughout this chapter that distributional parameters are known. For now we ignore the effects of Phase I estimations and assume that ZIP parameters p_0 and λ_0 are well known.

5.2.1 Solving charting constant L

For a given IC distribution with parameters p_0, λ_0 and w, the charting constant L is obtained according to the following consecutive steps:

1. We generate a data set of N runs, with in each run n observations. Let us denote the resulting observations with $\{Y_{j,1}, \ldots, Y_{j,n}\}$ for $j = 1, \ldots, N$. Each observation is ZIP(p_0, λ_0) distributed.
2. For all observations in each run, we compute the charting statistics E_n as defined in (4.3) for a specified value of w. Let us denote these resulting runs with $R_j = \{E_{j,1}, \ldots, E_{j,n}\}$ for $j = 1, \ldots, N$.
3. For an arbitrary value of L, we construct the ZIP-EWMA chart as defined in (5.1), and according to the specified values for p_0, λ_0 and w.
4. The run length of each run R_j in the ZIP-EWMA chart is determined and denoted with RL_j, for $j = 1, \ldots, N$. The ARL_0 is determined as the average of all computed run lengths, i.e. $ARL_0 = (RL_1 + \cdots + RL_N)/N$.
5. If the obtained ARL_0 does not equal the prespecified ARL_0, then we adjust L and repeat step 3-5.

Hence, the ARL_0 computations are based on N runs that each include n simulated observations. The implementation of these simulations in `R` is attached in Appendix C.1.

5.2.2 Choice of simulation size

The design of ARL simulations is based on a trade off between precision and computation time. In order to decrease computational time, a design choice was made to generate all observations at once as a $N \times n$ matrix. In order to generate reliable results, it is important to carefully chose the values of N and n. Alevizakos and Koukouvinos (2020) is used as reference study to select the number of iterations. Here, ARL simulations with similar performances are executed with the total amount of runs equal to 10.000. It is argued that 10,000 replications is enough to generate reliable results, since the research of Schaffer and Kim (2007) indicates that 5000 replications is enough for a standard EWMA chart. Therefore, it was also chosen in this study to fix the number of replications at 10,000 as well.

In addition, n should be large enough to ensure a very small probability of having no OC signal in the entire run. Hence, it is important to analyse the variance of the run lengths while choosing a value for n. However, the probability distribution of the run length is complex and therefore it difficult to find a closed form definition for the variance. Instead, simulations are executed to find a proper value for n. Here, test runs are simulated for various values of n, p_0, λ_0 and w to evaluate how many runs result in no OC signal. Candidate values of n are set at 2, 5 an 10 times the predefined ARL_0, since a higher IC average run length requires a larger simulation size. Values for the predefined ARL_0 are set at $200, 370$ and 500. The results are shown in Table B.1.

It is observed that the proportion of runs that does not return an OC signal less or equal to 0.01%, when choosing n as 10 times the predefined ARL_0 value. Hence, simulation size n is fixed at 2000, 3700 and 5000 when solving L for predefined ARL_0 values of 200, 370 an 500, respectively. The ZIP-EWMA control limits for various parameter settings are provided in Section 5.2.3. After this, the performance is evaluated by calculating the ARL_1 in Section 5.3.

5.2.3 IC control limits of ZIP-EWMA chart

We explore how the ZIP-EWMA performance is affected by different distributional parameters of the IC process. Hence, the ZIP-EWMA chart is constructed for IC scenarios where $p_0 = 0.3, 0.5, 0.8$ and $\lambda_0 = 3, 4$. These parameter values correspond to a ZIP distribution with multiple proportions of zero-inflation and a low expected value, which is common for monitoring data from high-purity processes in practice. In addition, we choose weight parameter $w = 0.2, 0.3$ and solve charting constant L to achieve an ARL_0 of 200, 370 and 500. The results for charting constant L are provided in Table 5.1

5.3 OC performance evaluation of the ZIP-EWMA chart

As stated in Section 5.1, the null hypothesis of the ZIP-EWMA chart assumes that all observations $\{Y_1, Y_2, \dots\}$ follow a ZIP(p_0, λ_0) distribution when the process is IC. The alternative hypothesis states that the process becomes OC after time $\mathcal{T}$, where at least one ZIP parameters changes. Hence, $Y_i \sim$ ZIP(p_1, λ_1) for $i = \mathcal{T}+1, \mathcal{T}+2, \dots$. Here, $p_1 \neq p_0$ and or $\lambda_1 \neq \lambda_0$. It is desired that the ZIP-EWMA chart produces an OC signal as soon as possible after the process becomes OC. Hence, we simulate the ARL_1 in order to evaluate the OC performance of the chart. Simulated ARL_1 values are obtained by execution of the following consecutive steps:

1. We assume that the IC process observations follow a ZIP distribution with known parameters p_0 and λ_0. A weight parameter w is defined and charting constant L is chosen from Table 5.1. The ZIP-EWMA control chart is constructed by substituting p_0, λ_0, w and L in (5.1).
2. We assume that the process becomes OC at the first time point, i.e., $\mathcal{T} = 0$. To evaluate OC performance, a data set is generated with N runs that each contain n observations. Let us denote these resulting runs with $\{Y_{j,\mathcal{T}+1}, \dots, Y_{j,\mathcal{T}+n}\}$ for $j = 1, \dots, N$. Each observation is ZIP(p_1, λ_1) distributed where $p_1 \neq p_0$ and or $\lambda_1 \neq \lambda_0$.

			Solved value L		
p_0	λ_0	w	$ARL_0 = 200$	$ARL_0 = 370$	$ARL_0 = 500$
0.3	3	0.2	2.5718	2.8312	2.9683
0.3	3	0.3	2.6883	2.9689	3.0962
0.3	4	0.2	2.5421	2.7609	2.8757
0.3	4	0.3	2.5848	2.8398	2.9546
0.5	3	0.2	2.6915	3.0098	3.1619
0.5	3	0.3	2.8699	3.1668	3.3103
0.5	4	0.2	2.6225	2.9194	3.0525
0.5	4	0.3	2.7583	3.0385	3.1657
0.8	3	0.2	3.1922	3.6200	3.8229
0.8	3	0.3	3.5349	3.9603	4.1537
0.8	4	0.2	3.1203	3.5291	3.7035
0.8	4	0.3	3.4247	3.8267	4.0068

Table 5.1: Solutions for constant L with $ARL_0 = 200, 370, 500$, $p_0 = 0.3, 0.5, 0.8$, $\lambda_0 = 3, 4$ and $w = 0.2, 0.3$.

3. For each run, and for all observations n, we compute the charting statistics E_n as defined in (4.3) and specified value for w. Let us denote these resulting runs with $R_j = \{E_{j,\mathcal{T}+1}, \ldots, E_{j,\mathcal{T}+n}\}$ for $j = 1, \ldots, N$.

4. The run length of each run R_j in the ZIP-EWMA chart is determined and denoted with RL_j, for $j = 1, \ldots, N$. The ARL_1 is determined as the average of all computed run lengths, i.e. $ARL_1 = (RL_1 + \cdots + RL_N)/N$.

Here, simulation size parameters N and n are chosen similarly as discussed in Section 5.2.2. Hence, $N = 10,000$ and n equals 10 times the predefined ARL_0 for which the ZIP-EWMA is designed. It is assumed in Step 1 that IC distributional parameters p_0 and λ_0 are known. Hence, the obtained ARL_1 reflects the baseline performance of the ZIP-EWMA chart.

In this chapter we consider OC scenarios that cause worse overall process performance, i.e. higher rate of defective pellets. The expected value of a ZIP(p,λ) distribution increases with λ and decreases with p. Therefore, more defects are observed in an OC scenario where λ_1 is larger than λ_0. Also, when p_1 is smaller than p_0. Hence, we analyse the ARL_1 for various OC scenarios where $\lambda_1 = \lambda + \delta_\lambda$ with $\delta_\lambda = 0.5, 1.0, 2.0$ and $p_1 = p_0 + \delta_p$ with $\delta_p = -0.1, -0.2, -0.3$. In order to compare the OC performance of the ZIP-EWMA chart for various proportions of zero-inflation, we evaluate the ARL_1 for three distinct IC distributions. The OC performance evaluation is executed for fixed parameters $\lambda_0 = 3$ and $w = 0.2$. Results for IC scenarios with $p_0 = 0.3, 0.5, 0.8$ are shown in Table 5.2, 5.3 and 5.4 respectively.

The results from Table 5.2, 5.3 and 5.4 show that all ARL_1 values of the ZIP-EWMA chart are significantly lower than the predefined ARL_0 values. This indicates that all OC distributional shifts are detected by the ZIP-EWMA chart. When comparing the ARL_1 results, it is observed that a distributional shift from (p_0, λ_0) to (p_1, λ_0) is detected faster when the IC distribution is more skewed, i.e., for a higher value of IC parameter $p_0 = 0.5, 0.8$ in Table 5.3 and 5.4. However, shift from (p_0, λ_0) to (p_0, λ_1) is detected faster in case the IC distribution is less skewed, i.e., for a lower value of IC parameter $p_0 = 0.3, 0.5$ in Table 5.2 and 5.3. An OC scenario in which both parameters (p_0, λ_0) change to (p_1, λ_1) is slightly detected faster in case IC parameter p_0 is small, i.e., for $p_0 = 0.3$ in Table 5.2.

		ARL_1 **value for:**		
OC scenario		$ARL_0 = 200$	$ARL_0 = 370$	$ARL_0 = 500$
p_1	λ_1	$L = 2.5718$	$L = 2.8312$	$L = 2.9683$
0.3	3.5	57.84	86.49	107.71
0.3	4.0	25.98	34.54	40.59
0.3	5.0	10.10	12.06	13.26
0.2	3.0	104.61	174.07	231.51
0.1	3.0	58.16	93.99	118.10
0.0	3.0	35.33	53.54	65.44
0.2	3.5	33.58	46.81	56.68
0.1	4.0	11.31	14.22	15.89
0.0	5.0	4.34	5.02	5.40

Table 5.2: ARL_1 values for all OC scenarios with IC parameters $p_0 = 0.3, \lambda_0 = 3$, $w = 0.2$, and $ARL_0 = 200, 370, 500$.

		ARL_1 **value for:**		
OC scenario		$ARL_0 = 200$	$ARL_0 = 370$	$ARL_0 = 500$
p_1	λ_1	$L = 2.6915$	$L = 3.0098$	$L = 3.1619$
0.5	3.5	69.10	107.81	140.94
0.5	4.0	34.69	48.40	58.54
0.5	5.0	14.09	17.72	20.28
0.4	3.0	91.15	159.72	209.00
0.3	3.0	49.01	78.12	100.15
0.2	3.0	29.73	43.97	54.72
0.4	3.5	35.93	54.05	64.93
0.3	4.0	12.85	16.46	18.89
0.2	5.0	5.16	6.07	6.52

Table 5.3: ARL_1 values for all OC scenarios with IC parameters $p_0 = 0.5, \lambda_0 = 3$, $w = 0.2$, and $ARL_0 = 200, 370, 500$.

5.4 Conclusion and discussion of the ZIP-EWMA chart

Results in Section 5.3 confirm that the ZIP-EWMA chart is able to detect when the process conditions are deteriorating. The control limits are determined according to predefined ARL_0 values and ARL_1 values are simulated for various OC distributional shifts. Table 5.2, 5.3 and 5.4 show for any distributional shift, as defined in the alternative hypothesis, that the ARL_1 is significantly lower than the predefined ARL_0 value. However, a higher p_0 value leads to faster detection of an OC distributional shift from (p_0, λ_0) to (p_1, λ_0), while a lower p_0 value leads to faster detection when (p_0, λ_0) shifts to (p_0, λ_1).

Based on these results, it is concluded that the ZIP-EWMA control chart could be used at Dow to monitor the high-purity processes of which the defect rate is not affected by any covariate. However, other SPC control charts should also be considered for performance evaluation and comparison when designing a monitoring scheme. In addition, the calculation of ARL values can also be further explored. Numerical approaches for exactly solving the ARL are not considered in this study, but they should be taken into account in further research. Finally, it should be mentioned that the weight parameter w of the ZIP-EWMA chart is not thoroughly researched in this study. Even though the control limits were defined

		ARL_1 **value for:**		
OC scenario		$ARL_0 = 200$	$ARL_0 = 370$	$ARL_0 = 500$
p_1	λ_1	$L = 3.1922$	$L = 3.6200$	$L = 3.8229$
0.8	3.5	97.32	158.46	204.06
0.8	4.0	57.59	86.27	106.12
0.8	5.0	28.27	37.36	43.29
0.7	3.0	67.12	110.98	144.05
0.6	3.0	31.99	48.87	59.46
0.5	3.0	19.18	27.13	32.33
0.7	3.5	37.36	56.31	65.71
0.6	4.0	13.74	18.00	20.23
0.5	5.0	5.89	7.17	7.73

Table 5.4: ARL_1 values for all OC scenarios with IC parameters $p_0 = 0.8, \lambda_0 = 3$, $w = 0.2$, and $ARL_0 = 200, 370, 500$.

for various values of w, ARL_1 computations are limited to $w = 0.2$. When constructing a ZIP-EWMA chart at Dow, it is recommended that a more detailed study is conducted to obtain the optimal value of w.

5.5 Summary

In this chapter, an EWMA control chart is proposed for data that follows an independent and identical ZIP distribution. Control limits are solved in Section 5.2, to obtain ARL_0 values of 200, 370 and 500, and for various IC parameters p_0, λ_0 and w. The OC performance of the ZIP-EWMA is evaluated in Section 5.3, for various distributional shifts. The ARL_1 results confirm that the ZIP-EWMA chart is able to detect when the process conditions are deteriorating. We keep the conclusions from Section 5.4 in mind while preceding to the following chapters, where we will focus on monitoring high-purity count data that is affected by a covariate.

6 Regression models for count data

In this chapter we consider high-purity count data that is affected by one covariate. It is explained in Chapter 2 that we consider a monitoring scheme for plastic pellet production, in which the observed number of defective pellets is affected by the inspected weight. The total amount of detected defects is denoted with Y_i, and the inspected weights are denoted with X_i for each point in time $i = 1, 2, \ldots$.

When monitoring the total number of detected defects over time, a correction must be made for the variable inspected weight. It is discussed in the literature review of Chapter 3 that one method to achieve this is with regression-based control charts. We can model the relationship between response variable Y_i and X_i with the zero-inflated regression models from Lambert (1992) and Heilbron (1994) for ZIP and ZINB distributed observations, respectively. These regression models are constructed according to the structure generalized linear models (GLM), while taking an additional proportion of zero inflation into account. For this reason, a certain level of understanding in GLM-theory is required before we can fully understand the structure of zero-inflated regression models. Hence, we focus on the theory regarding generalized linear models first in Section 6.1. Definitions of the ZIP and ZINB regression models are provided afterwards in Section 6.2.

6.1 Generalized linear models

It is described in Section 3.2 that generalized linear models are appropriate for modelling data of which the response variable follows a distribution from the EDM family. The EDM family is therefore discussed in the following section, after which the definition of a GLM is provided in Section 6.1. Estimation of regression coefficients and GLM residuals are discussed in Section 6.1.3 and 6.1.4, respectively. The textbook of Dunn and Smyth (2018) is used as a reference throughout each section. We consider general GLM theory where any random variable Y is affected by p covariates $X_1, \ldots, X_p$, in order to keep this section generic. In the context of plastic pellet production we have only one covariate, i.e. $p = 1$.

6.1.1 Exponential dispersion models

According to Dunn and Smyth (2018) (Section 5.3), a distribution for any random variable Y belongs to the EDM family if the probability function can be written in a specific form, defined in (6.1). This probability function is the probability density function if Y is continuous, and the probability mass function if Y is discrete. Hence, a probability distribution belongs to the EDM family if it can be written as

$$\mathcal{P}(y; \theta, \varphi) = a(y, \varphi) \exp\left\{ \frac{y\theta - \kappa(\theta)}{\varphi} \right\} \tag{6.1}$$

where,

- θ is called the canonical parameter.
- $\kappa(\theta)$ is a known function of the canonical parameter, called the cumulant function.
- $\varphi > 0$ is the dispersion parameter.
- $a(y, \varphi)$ is a normalising function ensuring that $\int_S \mathcal{P}(y; \theta, \varphi)dy = 1$ in case Y is continuous, and $\sum_{y \in S} \mathcal{P}(y; \theta, \varphi) = 1$ if Y is discrete, where S denotes the support of Y. The function $a(y, \varphi)$ does not necessarily have a closed form.

EDM distributions have some particular properties that are shared among all distributions in the family. First of all, the mean μ is a known one-to-one function of the canonical parameter θ. Therefore, it is common to use the notation $Y \sim EDM(\mu, \varphi)$, when stating that Y follows a distribution from the EDM family with mean μ and dispersion parameter φ. The probability function is also sometimes denoted as $\mathcal{P}(y; \mu, \varphi)$ for this reason. In addition, the variance of every EDM distribution is a function of the dispersion parameter and its expectation. Namely, the variance function is defined as $\text{Var}(Y) = \varphi V(\mu)$.

GLM theory relies heavily on the EDM probability function structure, since convergence of regression parameter estimation is proved for distribution of the form (6.1) only. This is further discussed in Section 6.1.3. As an example of how the EDM structure relates to well-known distributions, we can write the Poisson distribution in EDM format by defining the probability function as

$$\mathcal{P}(y; \mu, 1) = \exp\{y \log \mu - \mu - \log(y!)\} = \frac{\exp\{-\mu\}\mu^y}{y!}$$

by substituting (6.1) with $\theta = \log \mu$ as the canonical parameter, $\kappa(\theta) = \mu$ as the cumulant function, $\varphi = 1$ as the dispersion parameter and $a(y, \varphi) = 1/y!$. The variance function of a Poisson distribution is defined as $\text{Var}(Y) = 1 \cdot V(\mu) = \mu$. For each distribution in the EDM family, the variance function, canonical parameter, canonical function and dispersion parameter are uniquely defined. These definitions can be found for the most common EDMs in Dunn and Smyth (2018) (p.221). In the special case of a normal distribution we have constant variance, i.e., $V(\mu) = 1$ and $\varphi = \sigma^2$. Now that we have a general notation of all distributions in the EDM family, we can continue to define generalized linear models in the next section.

6.1.2 Definition of a Generalized Linear Model

Generalized linear model consists of two components. The first component is referred to as the random component, which models the distribution of response variable Y_i for $i = 1, 2, \ldots$. This distribution is denoted as $EDM(\mu_i, \varphi)$, with expected value μ_i and dispersion parameter φ. The random component is occasionally also defined as $Y_i \sim EDM(\mu_i, \varphi/w_i)$, where parameters w_i are used as non-negative weights to indicate the importance of each observation differently.

The second component is often referred to as the systematic component, which models the relation between response variable Y_i and all covariates $X_i = \{X_{i,1}, \ldots X_{i,p}\}$. The systematic component is defined as $g(\mu_i) = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}$, where $g(\cdot)$ is a known, monotonic, differentiable link function. The right-hand side of the systematic component is often referred to as the linear predictor, since it is a linear combination of all covariates. Parameters o_i are the offset for each individual observation, which are often set equal to zero. Regression coefficients $\beta = \{\beta_0, \ldots, \beta_p\}$ can be estimated for a given data set. Finally, the random and systematic component together define a generalized linear model as

$$\begin{cases} y_i \sim EDM(\mu_i, \varphi) \\ g(\mu_i) = o_i + \beta_0 + \sum_{j=1}^{p} \beta_j X_{i,j}. \end{cases} \tag{6.2}$$

The core structure of a GLM is specified by the choice of EDM distribution and the choice of link function. The canonical link function is defined such that $g(\mu) = \theta$, where θ is the canonical parameter from the EDM distribution. This link function is a common choice in practice since it ensures desirable statistical properties which generally lead to faster convergence of regression coefficient estimation. More specifically, the canonical link ensures that a minimal sufficient statistic exists for $\beta = \{\beta_0, \ldots, \beta_p\}$ (see e.g. McCullagh and Nelder (2019) (Section 2.2.4)). However, it is also possible to use alternative link functions as long as distributional properties are preserved. For example, $g(\mu_i) = \log(\mu_i)$ is the canonical link function of the Poisson distribution, since its canonical parameter is $\theta_i = \log(\mu_i)$. This link function ensures that μ_i is non-negative. However, the link function $g(\mu_i) = \sqrt{\mu_i}$ is also a common choice for Poisson regression, since it ensures $\mu_i \geq 0$ as well.

For a given data set, one can define a GLM by selecting an EDM distribution and a corresponding link function. Then, the GLM can be fitted to the data by estimating the regression coefficients. The exact estimation procedure is discussed in the following section.

6.1.3 Estimating regression coefficients

Let us define a data set $(y, x_1, \ldots, x_p) = \{(y_1, x_{1,1}, \ldots, x_{1,p}), \ldots, (y_n, x_{n,1}, \ldots, x_{n,p})\}$, with n observation and p covariates. Given that the observations follow a distribution form the EDM family, a GLM with link function $g(\cdot)$ can be fitted to estimate regression coefficients $\beta = \{\beta_0, \ldots, \beta_p\}$. The estimates are obtained by maximising the joint probability function of y, which is also referred to as the likelihood function (see e.g. Dunn and Smyth (2018) (Section 6.2)). When denoting the EDM probability function of random variable Y_i at point y_i with $\mathcal{P}(y_i; \mu_i, \varphi)$, as described in Section 6.1.1, then the likelihood function is defined as

$$\mathcal{L}(\beta_0, \ldots, \beta_p, \varphi; y) = \prod_{i=1}^{n} \mathcal{P}(y_i; \mu_i, \varphi)$$

where $\mu_i = g^{-1}(o_i + \beta_0 + \sum_{j=1}^{p} \beta_j x_{i,j})$. Maximisation of this likelihood function with respect to β provides the maximum likelihood (ML) estimates of regression coefficients, that are denoted with $\hat{\beta} = \{\hat{\beta}_0, \ldots, \hat{\beta}_p\}$. The logarithmic function is a monotonically increasing function. Hence, a maximum in the likelihood function is attained in the log-likelihood function as well. The log-likelihood function is however usually more convenient to work with, since it is defined by a sum instead of a product. Therefore, maximisation of the log-likelihood function used for GLM parameter estimation. The log-likelihood function is defined as

$$\ell(\beta_0, \ldots, \beta_p, \varphi; y) = \log \mathcal{L}(\beta_0, \ldots, \beta_p, \varphi; y) = \sum_{i=1}^{n} \log P(y_i; \mu_i, \varphi).$$

It is explained in Jørgensen (1997) (p.114) that maximisation of the log-likelihood is equivalent to minimisation of the total deviance function in case of an EDM. The total deviance function can be shown to be convex (see, e.g. Dunn and Smyth (2018) (Section 5.4.1)). Hence, a unique solution is obtained for ML estimates $\hat{\beta} = \{\hat{\beta}_0, \ldots, \hat{\beta}_p\}$. An additional advantage of EDM distributions, are their closed form definitions for the first and second derivative of the log-likelihood function. This assures that the iterative weighted least squares (IWLS) algorithm can be used to find the optimal solution. The details of this approach are thoroughly explained in Dunn and Smyth (2018) (Sections 6.2 and 6.3).

The fitted GLM is fully defined after estimating the regression coefficients, and obtaining maximum likelihood estimators $\hat{\beta} = \{\hat{\beta}_0, \ldots, \hat{\beta}_p\}$. The obtained model can then be used to predict the outcome of random variable Y_{n+1} based on covariates $x_{n+1,1}, \ldots, x_{n+1,p}$. The prediction accuracy depends on the goodness of fit of the GLM, which can be evaluated by inspection of regression residuals. Definitions of GLM residuals are provided and discussed in the following section.

6.1.4 GLM residuals

For a defined regression model, we can evaluate the residuals to assess the goodness of fit. Let us assume that we have a data set $(y, x_1, \ldots, x_p) = \{(y_1, x_{1,1}, \ldots, x_{1,p}), \ldots, (y_n, x_{n,1}, \ldots, x_{n,p})\}$, with n observation and p covariates. In addition, lets us denote the GLM with defined link function $g(\cdot)$, offsets o_i and ML regression coefficients $\hat{\beta} = \{\hat{\beta}_0, \ldots, \hat{\beta}_p\}$. Then, predictions for each observation in $y = (y_1, \ldots, y_n)$ are denoted with $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, whereas $\hat{\mu}_i = g^{-1}(o_i + \hat{\beta}_0 + \sum_{j=1}^{p} \hat{\beta}_j x_{i,j})$ for $i = 1, \ldots, n$. The distance $y_i - \hat{\mu}_i$ is often referred to as the raw residual.

Recall that the variance of any EDM distribution is a function of the expected value, i.e. $V(\mu_i)$. Hence, the variance of each observation y_i depends on the value of its expected value μ_i, and therefore we do not have identical distribution of observations. A direct consequence of this property is that the raw residuals also do not have identical distributions. More specifically, it is possible that $r_i = y_i - \hat{\mu}_i$ and $r_j = y_j - \hat{\mu}_j$ have different variances when $\mu_i \neq \mu_j$. In order to obtain identically distributed residuals that can be used for goodness of fit assessment, we need to correct for the inconstant variance. Dunn and Smyth (2018) (Chapter 8.3) describe three distinct methods for obtaining approximately identically distributed GLM residuals. Namely, Pearson residuals, deviance residuals and quantile residuals. Each method is briefly discussed in the following sections.

Pearson residuals

Pearson residuals are the most intuitive and direct approach of obtaining GLM residuals. Similarly to Pearson residuals in linear regression, the idea is to scale the raw residuals by dividing out their inconstant standard deviation. The estimated standard deviation of the raw residual $r_i = y_i - \hat{\mu}_i$ equals the estimated standard deviation of observation Y_i, since prediction $\hat{\mu}_i$ is a known value. An estimation of the standard deviation of Y_i it given by the square root of the variance function evaluated at point $\hat{\mu}_i$, i.e., $\sqrt{V(\hat{\mu}_i)}$. For a given set of observations $y = (y_1, \ldots, y_n)$ and predictions $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, Pearson residuals are defined in vector notation as

$$r^P = \frac{y - \hat{\mu}}{\sqrt{V(\hat{\mu})}} \tag{6.3}$$

where $V(\cdot)$ is the variance function of the EDM distribution, for which the GLM is constructed. These Pearson residuals are the square root of the unit Pearson statistic, which is approximately Chi-square distributed when the central limit theorem applies. This theorem applies under various conditions that are thoroughly explained in Dunn and Smyth (2018) (Section 7.5). Under the same conditions, Pearson residuals have an approximate normal distribution. However, it should be mentioned that Pearson residuals can be far from normal when the central limit theorem does not apply. This is often the case for discrete EDM distributions and especially for discrete distributions with low expected values (see e.g., Feller (1945) or Jolliffe (1995)). If the central limit theorem does not hold, then normality of Pearson residuals is unlikely.

Deviance residuals

A different distance measure between y_i and $\hat{\mu}_i$ is the unit deviance $d(y_i, \hat{\mu}_i)$. This unit deviance is twice the difference in log-likelihood between the saturated model and the fitted model, multiplied by dispersion parameter φ. This generalises to the residual sum of squares in ordinary linear regression. In GLM, the saturated model at point y_i is the EDM with expected value y_i, and the fitted model is the EDM with expected value $\hat{\mu}_i$. The unit deviance of a GLM is defined as $d(y, \mu) = 2t(y, y) - t(y, \mu)$ with $t(y, \mu) = y\theta + \kappa(\theta)$. Here, θ is the canonical parameter of the EDM distribution. Notice that $t(y, \mu)$ is expressed in terms of y and θ since μ and θ are one-to-one functions of each other. The overall measure of distance between y and $\hat{\mu}$ is provided by the deviance function that is defined as $D(y, \hat{\mu}) = \sum_{i=1}^{n} d(y_i, \hat{\mu}_i)$. This deviance function is equivalent to the sum of squared estimate of errors (SSE) in case of normally

distributed observations. Hence, we can consider the overall deviance to be a generalisation of the SSE to GLM regression. Deviance residuals are defined as the signed square root of the unit deviance. For a given set of observations $y = (y_1, \ldots, y_n)$ and predictions $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$, we can define deviance residuals as

$$r^D = \text{sign}(y - \hat{\mu})\sqrt{d(y, \hat{\mu})}. \tag{6.4}$$

The deviance statistic has an approximate Chi-square distribution, when the saddle point approximation applies to the EDM distribution. This approximation applies under various conditions that are again explained in Dunn and Smyth (2018) (Section 7.5). Under the same conditions, deviance residuals have an approximate normal distribution. However, deviance residuals can be far from normally distributed when the saddle point approximation does not apply. This is again often the case for discrete distributions with low expected values. Nevertheless, deviance residuals are more likely to be normally distributed than Pearson residuals, since the central limit theorem has a slower convergence rate then the saddle point approximation, i.e., $O(\varphi^{1/2})$ instead of $O(\varphi)$. This is explained in Dunn and Smyth (2018) (Section 7.5). To illustrate this, let us consider an example where observations $y = (y_1, \ldots, y_n)$ are collected from a Poisson distributed random variable. Then, the saddle point approximation is sufficiently accurate for $\min(y) \geq 3$, while the central limit theorem is only sufficiently accurate for individual observations when $\min(y) \geq 5$. This is explained in Dunn and Smyth (2018) (Section 7.5) as well. Hence, deviance residuals are more likely to be approximately normal than Pearson residuals in this case.

Finally, notice that the unit deviance is only defined for EDM distributions that allow $\mu = y$. Otherwise, the unit deviance is approximated by choosing y close to μ, which is explained in Dunn and Smyth (2018) (Section 5.4.1).

Quantile residuals

Dunn and Smyth (1996) introduced quantile residuals for GLM regression as an alternative to both Pearson and deviance residuals. Quantile residuals have exactly normal distribution apart from the sampling variability in estimating distribution parameters. The definition of quantile residuals is different for continuous and discrete GLMs. Let $F(y_i; \hat{\mu}_i, \varphi)$ be the cumulative distribution function of random variable Y at point y_i, with parameters $\hat{\mu}_i$ and φ. Also, lets define a given set of observations $y = (y_1, \ldots, y_n)$ with predictions $\hat{\mu} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$. If Y is a continuous random variable, quantile residuals is given by

$$r^Q = \Phi^{-1}\{F(y; \hat{\mu}, \varphi)\} \tag{6.5}$$

where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal distribution. If Y is a discrete random variable, then let $b = F(y; \hat{\mu}, \varphi)$ and $a = \lim_{\varepsilon \to 0^-} F(y + \varepsilon; \hat{\mu}, \varphi)$. Here, we use the left-hand limit such that $y + \varepsilon < y$. The quantile residuals in vector notation are given by

$$r^Q = \Phi^{-1}(U) \tag{6.6}$$

where U is a vector of uniform random variables on the interval $(a, b]$. We often refer to quantile residuals for discrete GLMs as randomised quantile residuals, because of the randomisation that is inserted with U. As mentioned before, quantile residuals and randomised quantile residuals are both continuous and normally distributed, apart from the sampling variability in estimating μ_i. This implies that quantile residuals are exactly normally distributed in case n is large and $\hat{\mu}_i = \mu_i$ for $i = 1, 2, \ldots$. A deviation from normality can be observed for less well-fitted models. A proof of this property is provided for the continuous and discrete case in Lemmas A.2.1 and A.2.2, respectively (see Appendix A).

This concludes the theoretical part regarding GLM. Now that we have discussed the definitions for the EDM family, GLM regression models, parameter estimation procedures and definitions for GLM residuals, we can proceed to discuss the regression models of interest in the following section.

6.2 Zero-inflated regression models

Regression models for zero-inflated count data originate from Lambert (1992), which introduced the zero-inflated Poisson (ZIP) model. A regression model for zero-inflated negative binomial (ZINB) response was introduced not much later by Heilbron (1994). Both zero-inflated distributions do not belong to the EDM family since they cannot be written in the specific form that is defined in (6.1). Zero-inflated regression models can therefore not be classified as GLM, even though both models show great resemblance with the GLM structure. Namely, the expected value of the response variable is modelled through a non-linear link function and a linear predictor that includes all covariates.

In the context of plastic pellet production, the zero-inflated regression models are employed to estimate the relationship between the total amount of detected defects Y_i, and inspected weights X_i for each point in time $i = 1, 2, \ldots$. Therefore, we focus on one-dimensional zero-inflated regression models, where one covariate is included. Definitions of the one-dimensional ZIP and ZINB regression model are provided in Sections 6.2.1 and 6.2.2, respectively, along with an explanation of parameter estimation methods and definitions of Pearson, deviance and randomised quantile residuals. The defined ZIP and ZINB model are employed in a regression-based Shewhart chart in Chapter 7.

6.2.1 Zero-inflated Poisson model

High-purity count data processes inherit a particularly large amount of zero observations, which is not accounted for by the Poisson distribution. Hence, modelling zero-inflated data with a Poisson GLM may cause violation of the equidispersion assumption, which leads to inaccurate estimates. As an alternative, Lambert (1992) introduced the zero-inflated Poisson (ZIP) model, of which a one-dimensional variant is formalised in (6.7). Here, the response variable Y_i is assumed to follow a ZIP(p_i,λ_i) distribution as defined in (2.1), where the effect of the inspected weight X_i is modelled through parameters p_i and λ_i. Hence, observations are not identically distributed in this model. With probability $1 - p_i$, variable Y_i follows a Poisson distribution with expected value λ_i. With probability p_i we have that Y_i equals a structural zero.

$$P(Y_i = y_i|X_i) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & \text{if } y_i = 0 \\ (1-p_i)\dfrac{e^{-\lambda_i}\lambda_i^{y_i}}{y_i!} & \text{if } y_i > 0 \end{cases} \tag{6.7}$$

$$\text{where} \quad \lambda_i = \exp(\beta_0 + \beta_1 X_i) \quad \text{and} \quad p_i = \frac{\exp(\gamma_0 + \gamma_1 X_i)}{1 + \exp(\gamma_0 + \gamma_1 X_i)}$$

It is explained in Section 3.2 that two variants of the ZIP regression model are defined by Lambert (1992). The ZIP(τ) model is defined for a situation where parameters p_i and λ_i depend on the same set of covariates. In this thesis, we consider a situation where only one covariate affects the response, such that definitions for p_i and λ_i in (6.7) depend on the same covariate X_i. Hence, a ZIP(τ) model is considered in this thesis, which we will abbreviate to ZIP in the following sections. The ZIP expected value, conditional on the value of X_i, is defined as

$$E[Y_i|X_i] = (1-p_i)\lambda_i \tag{6.8}$$

where the condition on X_i indicates that values for p_i and λ_i are known. Similarly, the conditional variance is defined as

$$\text{Var}(Y_i|X_i) = (1-p_i)(\lambda_i + p_i\lambda_i^2). \tag{6.9}$$

A proof of the ZIP expected value and variance is appended in Lemmas A.1.1 and A.1.2, respectively. The ZIP regression model is similar to a GLM because the the expected value $E[Y_i|X_i] = (1-p_i)\lambda_i = \mu_i$

is modelled through a linear predictor of the covariate, and a Log and Logit link functions for λ_i and p_i, respectively. The Log link function ensures that expected value λ_i is strictly positive and the Logit link function ensures that p_i is estimated between 0 and 1. This GLM structure that combines a linear predictor with a link function is explained in Section 6.1.2.

Let us consider a data set $(y, x) = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ with n observations. Maximum likelihood estimates for parameters β_0, β_1, γ_0 and γ_1 are obtained by maximising the ZIP log-likelihood function that is defined as

$$\ell(\beta_0, \beta_1, \gamma_0, \gamma_1; y) = \sum_{i:\, y_i=0} \log\left[\exp(\gamma_0 + \gamma_1 x_i) + \exp(-\exp(\beta_0 + \beta_1 x_i))\right] + \sum_{i:\, y_i>0} \left[(y_i - 1)\exp(\beta_0 + \beta_1 x_i) - \log(y_i!)\right] - \sum_{i=1}^{n} \log\left[1 + \exp(\gamma_0 + \gamma_1 x_i)\right].$$

The maximisation procedure is executed by means of the Broyden–Fletcher–Goldfarb –Shanno algorithm (BFGS). The parameter estimates are proved to have asymptotically normal distribution in Lambert (1992), where extensive simulations show that estimates can be trusted when the ZIP(τ) are fitted on sufficiently large data sets, i.e. $n \geq 100$. The obtained ML estimates for regression coefficients β_0, β_1, γ_0, and γ_1 are denoted with $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$, respectively. With these estimates, we can also obtain estimates for parameters $\lambda = (\lambda_1, \ldots, \lambda_n)$ and $p = (p_1, \ldots, p_n)$ by means of their definitions in (6.7). These estimates are denoted with $\hat{\lambda} = (\hat{\lambda}_1, \ldots, \hat{\lambda}_n)$ and $\hat{p} = (\hat{p}_1, \ldots, \hat{p}_n)$, respectively. Finally, a set of predictions for observations y are obtained with $\hat{\mu} = (1 - \hat{p})\hat{\lambda} = (\hat{\mu}_1, \ldots, \hat{\mu}_n)$.

We can evaluate the goodness of fit of the regression model by analysis of the residuals. It is explained in Section 6.1.4 that three different types of residuals can be obtained from a GLM. The same types of residuals can be obtained from the ZIP regression model. First of all, Pearson residuals computed similarly to GLM Pearson residuals, which are defined in Section 6.1.4. Hence, ZIP Pearson residuals are obtained by substituting the ZIP expected value from (6.8) and the ZIP variance from (6.9) into (6.3). The following definition for ZIP Pearson residuals is obtained

$$r_i^P = \frac{y_i - (1 - \hat{p}_i)\hat{\lambda}_i}{\sqrt{(1 - \hat{p}_i)(\hat{\lambda}_i + \hat{p}_i\hat{\lambda}_i^2)}} \tag{6.10}$$

for all observations $i = 1, \ldots, n$. Next, deviance residuals for zero-inflated regression models are less trivial to define. From Section 6.1.4 we know that in case of GLM regression, we can easily define the unit deviance as a function of EDM parameters. However, this is not possible for zero-inflated regression models for which those EDM parameters cannot be defined. Nevertheless, we can still define the unit deviance of zero-inflated models as the difference between the log-likelihood of the saturated and the fitted model. It is stated in Feng et al. (2020) that for a ZIP regression model at point y_i, the saturated model is defined as the Poisson regression model with expected value y_i. Hence, ZIP deviance residuals for $i = 1, \ldots, n$ are defined as $r_i^D = \text{sign}(y_i - \hat{\mu}_i)(2\{\log g_1(y_i|y_i) - \log f_1(y_i|\hat{p}_i, \hat{\lambda}_i)\})^{1/2}$, where $g_1(\cdot|\lambda)$ is the Poisson probability mass function with expected value λ and $f_1(\cdot|p, \lambda)$ is the ZIP probability mass function with parameters p and λ as defined in (2.1). Written in the extensive form, ZIP deviance residuals are therefore defined as

$$r_i^D = \text{sign}(y_i - \mu_i)\Big(2 \cdot \Big\{ - y_i + y_i \log y_i - \log(y_i!) - \mathbb{1}_{\{y_i=0\}} \log\left[\hat{p}_i + (1 - \hat{p}_i)e^{-\hat{\lambda}_i}\right] - \mathbb{1}_{\{y_i>0\}} \left[\log(1 - \hat{p}_i) - \hat{\lambda}_i + y_i \log \hat{\lambda}_i - \log(y_i!)\right] \Big\}\Big)^{1/2} \tag{6.11}$$

for observations $i = 1, \ldots, n$. The third and final residual type is the quantile residual. In this case,

we use randomised quantile residuals since the ZIP distribution is discrete. These residuals are defined in the exact same way as was explained for GLM in Section 6.1.4. Let us define the ZIP cumulative distribution function at point y_i as as $F_1(y_i; \hat{\lambda}_i, \hat{p}_i)$, with parameters $\hat{\lambda}_i$ and $\hat{p}_i$. Then, ZIP randomised quantile residual for observations $i = 1, \ldots, n$ are defined as

$$r_i^Q = \Phi(u_i) \tag{6.12}$$

where u_i is a uniform random variable between $a_i = \lim_{y \to y_i} F_1(y \,|\, \hat{p}_i, \hat{\lambda}_i)$ and $b_i = F_1(y_i \,|\, \hat{p}_i, \hat{\lambda}_i)$.

When constructing a regression-based control chart for ZIP distributed data, we can monitor Pearson, deviance and randomised quantile residuals. A ZIP regression-based Shewhart chart is defined for all three types of residuals in Chapter 7. Besides the ZIP model, zero-inflated count data can also be modelled by means of the ZINB distribution. The ZINB model is often considered as an alternative to the ZIP model since it includes a dispersion parameter that allows for adjustable amounts of variation in the data. The ZINB model is defined an discussed in the following section.

6.2.2 Zero-inflated negative binomial model

As a generalisation to the ZIP model, Heilbron (1994) introduced a zero-inflated negative binomial (ZINB) regression model. A one-dimensional variant of this regression model is formalised in (6.13). Here, the response variable Y_i is assumed to follow a ZINB(p_i,λ_i,τ) distribution as defined in (2.2), where the effect of the inspected weight X_i is again modelled through parameters p_i and λ_i. Hence, observations are not identically distributed in this model. The strictly positive size parameter is denoted with τ. With probability $1 - p_i$, variable Y_i follows a negative binomial distribution with expected value λ_i. With probability p_i we have that Y_i equals zero.

$$\begin{gathered} P(Y_i = y_i \,|\, X_i) = \begin{cases} p_i + (1-p_i)\left(1 + \frac{\lambda_i}{\tau}\right)^{-\tau} & \text{if } y_i = 0 \\ (1-p_i)\frac{\Gamma(y_i+\tau)}{y_i!\Gamma(\tau)}\left(1 + \frac{\lambda_i}{\tau}\right)^{-\tau}\left(1 + \frac{\tau}{\lambda_i}\right)^{-y_i} & \text{if } y_i > 0 \end{cases} \\ \text{where} \quad \lambda_i = exp(\beta_0 + \beta_1 X_i) \quad \text{and} \quad p_i = \frac{exp(\gamma_0 + \gamma_1 X_i)}{1 + exp(\gamma_0 + \gamma_1 X_i)} \end{gathered} \tag{6.13}$$

The ZINB expected value, conditional on the value of X_i, is defined as

$$E[Y_i|X_i] = (1 - p_i)\lambda_i \tag{6.14}$$

where the condition on X_i indicates again that values for p_i and λ_i are known. Similarly, the conditional variance is defined as

$$\text{Var}(Y_i|X_i) = \lambda_i(1 - p_i)(1 + p_i\lambda_i + \lambda_i/\tau). \tag{6.15}$$

A proof of the expectation and variance is appended in Lemmas A.1.3 and A.1.4, respectively. The Log and Logit link functions are again applied for the same reasons as discussed in Section 6.2.1.

Let us consider again a data set $(y, x) = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ with n observations. Maximum likelihood estimates for parameters β_0, β_1, γ_0,γ_1 and τ are obtained by maximising the ZINB log-likelihood

function that is defined as

$$\begin{aligned}\ell(\beta_0,\beta_1,\gamma_0,\gamma_1,\tau;y) = &\sum_{i=1}^{n}\log\left[1+\exp(\gamma_0+\gamma_1 x_i)\right]\\ &-\sum_{i:\,y_i=0}\log\left[\exp(\gamma_0+\gamma_1 x_i)+\left(\frac{\exp(\beta_0+\beta_1 x_i)+\tau}{\tau}\right)^{-\tau}\right]\\ &+\sum_{i:\,y_i>0}\log\left[\tau\left(\frac{\exp(\beta_0+\beta_1 x_i)+\tau}{\tau}\right)+y_i\log\left(1+\tau\exp(\beta_0+\beta_1 x_i)\right)\right]\\ &+\sum_{i:\,y_i>0}\log(\Gamma(\tau))+\log(\Gamma(1+y_i))-\log(\Gamma(\tau+y_i)).\end{aligned}$$

The maximisation procedure is executed by means of the Broyden–Fletcher–Goldfarb –Shanno algorithm (BFGS). It is proved by Heilbron (1994) that parameter estimates of ZINB regression have asymptotically normal distribution. The obtained ML estimates for regression coefficients β_0, β_1, γ_0, and γ_1 are denoted with $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$ and $\hat{\gamma}_1$, respectively. If the true value of τ is unknown, then the ML estimate is obtained which is denoted with $\hat{\tau}$. With these estimates, we can also obtain estimates for parameters $\lambda = (\lambda_1,\ldots,\lambda_n)$ and $p = (p_1,\ldots,p_n)$ by means of their definitions in (6.13). These estimates are denoted with $\hat{\lambda} = (\hat{\lambda}_1,\ldots,\hat{\lambda}_n)$ and $\hat{p} = (\hat{p}_1,\ldots,\hat{p}_n)$, respectively. Finally, a set of predictions for observations y are obtained with $\hat{\mu} = (1-\hat{p})\hat{\lambda} = (\hat{\mu}_1,\ldots,\hat{\mu}_n)$.

Once the ZINB model is fitted and estimates for all parameters are obtained, we can evaluate the goodness of fit by analysis of residuals. Pearson residuals for ZINB regression computed similarly to GLM Pearson residuals, which are defined in Section 6.1.4. Hence, ZINB Pearson residuals are obtained by substituting the ZINB expected value from (6.14) and the ZIP variance from (6.15) into (6.3). The following definition of Pearson residuals for ZINB regression is obtained

$$r_i^P = \frac{y_i-(1-\hat{p}_i)\hat{\lambda}_i}{\sqrt{\hat{\lambda}_i(1-\hat{p}_i)(1+\hat{p}_i\hat{\lambda}_i+\hat{\lambda}_i/\hat{\tau})}} \tag{6.16}$$

for all observations $i = 1,\ldots,n$. It was already explained in Section 6.2.1 that the unit deviance of zero-inflated models is defined as the difference between the log-likelihood of the saturated and the fitted model. Hence, deviance residuals are defined as the signed squared root of the unit deviance. It is stated in Feng et al. (2020) that for a ZINB regression model at point y_i, the saturated model is defined as the negative binomial regression model with expected value y_i and size parameter τ. If the true value of τ is unknown, then the ML estimate $\hat{\tau}$ can be applied. Hence, ZINB deviance residuals for $i = 1,2,\ldots$ are defined as $r_i^D = \operatorname{sign}(y_i-\hat{\mu}_i)(2\{\log g_2(y_i|y_i,\hat{\tau}) - \log f_2(y_i|\hat{p}_i,\hat{\lambda}_i,\hat{\tau})\})^{1/2}$ where $g_2(\cdot|\lambda,\tau)$ is the negative binomial probability mass function with expected value λ and size parameter τ, and $f_2(\cdot|p,\lambda,\tau)$ is the ZINB probability mass function with parameters p,λ and τ as defined in (2.2). Written in the extensive form, ZINB deviance residuals are therefore defined as

$$\begin{aligned}r_i^D = \operatorname{sign}(y_i-\mu_i)\Bigg(2\cdot\Bigg\{&\log\frac{\Gamma(y_i+\hat{\tau})}{\Gamma(\hat{\tau})\Gamma(y_i+1)}+y_i\log\left(\frac{y_i}{y_i+\hat{\tau}}\right)+\hat{\tau}\log\left(\frac{\hat{\tau}}{y_i+\hat{\tau}}\right)\\ &-\mathbb{1}_{\{y_i=0\}}\log\left[\hat{p}_i+(1-\hat{p}_i)\left(\frac{\hat{\tau}}{\hat{\lambda}_i+\hat{\tau}}\right)^{\hat{\tau}}\right]\\ &-\mathbb{1}_{\{y_i>0\}}\left[\log(1-\hat{p}_i)+\log\frac{\Gamma(y_i+\hat{\tau})}{\Gamma(\hat{\tau})\Gamma(y_i+1)}+y_i\log\left(\frac{y_i}{y_i+\hat{\tau}}\right)+\hat{\tau}\log\left(\frac{\hat{\tau}}{y_i+\hat{\tau}}\right)\right]\Bigg\}\Bigg)^{1/2}\end{aligned} \tag{6.17}$$

for observations $i = 1,\ldots,n$. The third and final residual type is the quantile residual. In this case, we use randomised quantile residuals since the ZINB distribution is discrete. These residuals are defined

in the exact same way as was explained for GLM in Section 6.1.4. Let us define the ZINB cumulative distribution function at point y_i as as $F_2(y_i; \hat{\lambda}_i, \hat{p}_i, \hat{\tau})$, with parameters $\hat{\lambda}_i$, $\hat{p}_i$ and τ. Then, ZIP randomised quantile residual for observations $i = 1, \ldots, n$ are defined as

$$r_i^Q = \Phi(u_i) \tag{6.18}$$

where u_i is a uniform random variable between $a_i = \lim_{y \to y_i} F_2(y \,|\, \hat{p}_i, \hat{\lambda}_i, \hat{\tau})$ and $b_i = F_2(y_i \,|\, \hat{p}_i, \hat{\lambda}_i, \hat{\tau})$. When constructing a regression-based control chart for ZIP distributed data, we can again monitor Pearson, deviance and randomised quantile residuals. A ZINB regression-based Shewhart chart is defined for all three types of residuals in Chapter 7.

When dealing with zero-inflated data in practice, it is important to carefully decide which regression model to use. The ZINB model is often considered as an alternative to the ZIP model, since it includes a size parameter τ. This size parameter allows for adjustable amounts of variation in the data, whereas the ZINB variance of the negative binomial distribution decreases when τ increases. The ZIP model on the other hand is more simplistic to interpret. In this research, we do not work with real plant data such that it is not necessary to select one model in particular. Nevertheless, it should be mentioned that a model selection procedure is provided in Mahmood (2020), which describes how we can choose between a Poisson, ZIP, negative binomial or ZINB model.

Here, a likelihood-ratio-test (LRT) is proposed to test whether the data contains overdispersion with respect to a Poisson distribution. The hypotheses of the LRT test are defined as $H_0 : \tau = 0$ or $H_1 : \tau > 0$. Hence, we should use a negative binomial distribution in case the null hypothesis is rejected. In addition to the LRT test, a Vuong test for zero-inflation is proposed to compare the traditional Poisson model with the ZIP model, and the negative binomial model with the ZINB model. This is a log-likelihood ratio test that originates from Vuong (1989), where a test statistic V is proposed that follows an asymptotic standard normal distribution. The zero-inflated model is preferred over the traditional model when $|V| < \Phi^{-1}(1 - \alpha/2)$, where α is the significance level and $\Phi^{-1}(\cdot)$ is the inverse standard normal cdf. When working with real data, it is important to follow these steps of model selection to determine which model is the best fit. In addition to the LRT-Vuong method, Xie et al. (2001) provides an overview of alternative statistical tests that can be used to determine whether a zero-inflated model should be applied.

6.3 Summary

In this thesis, we focus on ZIP and ZINB regression-based Shewhart charts. Both the ZIP and ZINB regression model are defined according to the structure of generalized linear models, such that we start with revising the GLM theory. A definition of exponential dispersion models is provided in Section 6.1.1, after which generalized linear models are defined in Section 6.1.2. The maximum likelihood procedure for estimating regression coefficients is described in Section 6.1.3, which is followed by a definition of GLM residuals in Section 6.1.4. Here, Pearson, deviance and quantile residuals are defined. It is discussed that quantile residuals have exact standard normal distribution under a regression model with perfect estimates for the regression coefficients. Pearson residuals have approximate normal distribution when the central limit theorem holds for individual observations. Deviance residuals have approximate normal distribution when the saddle point approximation applies.

The zero-inflated regression models from Lambert (1992) and Heilbron (1994) are defined in Sections 6.2.1 and 6.2.2, respectively. Estimation of regression coefficients is described in both sections after which Pearson, deviance and randomised quantile residuals are defined for each model. The construction of ZIP and ZINB regression-based Shewhart charts is discussed in the next chapter.

7 | Monitoring count data with covariates

In this chapter, we design a regression-based monitoring scheme for high-purity count data that is affected by one covariate. More specifically, we construct a ZIP and ZINB regression-based Shewhart chart for predictive Pearson, deviance and randomised quantile residuals. Regression models (6.7) and (6.13) are used throughout this chapter. A definition of the ZIP and ZINB regression-based Shewhart chart is provided in Section 7.1, along with the hypotheses of the chart. Simulation of zero-inflated data and distribution of ZIP and ZINB regression residuals is discussed in Sections 7.2 and 7.3, respectively. It is described in Section 7.4 that we follow two strategies for construction and performance evaluation of each chart. Finally, the construction and OC performance evaluation procedures are discussed in Sections 7.5 and 7.6.

7.1 Regression-based Shewhart charts

In the context of plastic pellet production at Dow, let us denote the total number of defective pellets with Y_i and the inspected weights with X_i at time $i = 1, 2, \ldots$. When monitoring the amount of detected defective pellets Y_i over time, we must correct for the inspected weights X_i. Regression-based control charts can be employed for this purpose. Let us define a data set $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$ of size m, that reflects the IC process. It was described in Section 4.1 that we refer to this data as the Phase I data. The construction of a regression-based Shewhart chart is discussed in the following sections, for ZIP and ZINB distributed observations Y_i, respectively.

7.1.1 The ZIP regression-based Shewhart chart

Let us assume that random a variable Y_i follows a ZIP distribution. Then, we can fit the one dimensional ZIP regression model that is defined in (6.7), to the Phase I data (y, x) in order to obtain estimates for regression coefficients $\beta_0, \beta_1, \gamma_0$ and γ_1. Let us denote these estimates with $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0$ and $\hat{\gamma}_1$, respectively. Now let (y_{m+i}, x_{m+i}) be a new observation with $i = 1, 2, \ldots$, where m denotes the size of the Phase I data. We can use the fitted ZIP model to predict the value of y_{m+i}, based on the value of x_{m+i}. Let us denote the prediction of y_{m+i} with $\hat{\mu}_{m+i}$. It is described in Section 6.2.1 that

$$\hat{\mu}_{m+i} = (1 - \hat{p}_{m+i})\hat{\lambda}_{m+i} \tag{7.1}$$

where

$$\hat{p}_{m+i} = \frac{\exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+i})}{1 + \exp(\hat{\gamma}_0 + \hat{\gamma}_1 x_{m+i})} \tag{7.2}$$

and

$$\hat{\lambda}_{m+i} = \exp(\hat{\beta}_0 + \hat{\beta}_1 x_{m+i}). \tag{7.3}$$

Here, $\hat{\mu}_{m+i}$ represents the ZIP expected value that is defined in (6.8). For each new observation (y_{m+i}, x_{m+i}) with $i = 1, 2, \dots$ and corresponding predictions $\hat{\mu}_{m+i}, \hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$, we can obtain regression residuals. Namely, ZIP Pearson residuals are obtained by substituting $y_{m+i}, \hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$ into (6.10). Let us denote these residuals with $r^P = \{r^P_{m+1}, r^P_{m+2}, \dots\}$. Similarly, ZIP deviance residuals are obtained by substituting $y_{m+i}, \hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$ into (6.11). Let us denote the ZIP deviance residuals with $r^D = \{r^D_{m+1}, r^D_{m+2}, \dots\}$. Finally, ZIP randomised quantile residuals are denoted with $r^Q = \{r^Q_{m+1}, r^Q_{m+2}, \dots\}$ and obtained by substituting $y_{m+i}, \hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$ into (6.12). In a regression-based Shewhart chart, we can monitor Pearson, deviance residuals or randomised quantile residuals over time.

As long as the process is in control, it is assumed that observations Y_{m+i} with $i = 1, 2, \dots$ follow a ZIP distribution with parameters $\hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$. These parameters are predicted based on the observed value of X_{m+i} and established regression coefficients $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0$ and $\hat{\gamma}_1$, as defined in (7.2) and (7.3), respectively. Let us denote this relation with $Y_{m+i} \sim \text{ZIP}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}|X_{m+i})$. When monitoring ZIP regression residuals over time in an SPC control chart, we can define the hypotheses of this chart as

$$
\begin{aligned}
H_0 :&\quad Y_{m+i} \sim \text{ZIP}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}|X_{m+i}) \quad \text{for} \quad i = 1, 2, \dots \\
H_1 :&\begin{cases} Y_{m+i} \sim \text{ZIP}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}|X_{m+i}) & \text{for} \quad i = 1, \dots, \mathcal{T} \\ Y_{m+i} \sim \text{ZIP}(p^{OC}_{m+i}, \lambda^{OC}_{m+i}|X_{m+i}) & \text{for} \quad i = \mathcal{T}+1, \mathcal{T}+2, \dots . \end{cases}
\end{aligned} \tag{7.4}
$$

where $\hat{p}_{m+i} \neq p^{OC}_{m+i}$ and or $\hat{\lambda}_{m+i} \neq \lambda^{OC}_{m+i}$. Hence, the process becomes out of control after change point $\mathcal{T}$. For a known observation (y_{m+i}, x_{m+i}), parameters p^{OC}_{m+i} and λ^{OC}_{m+i} are defined as

$$
p^{OC}_{m+i} = \frac{\exp(\gamma_0^{OC} + \gamma_1^{OC} x_{m+i})}{1 + \exp(\gamma_0^{OC} + \gamma_1^{OC} x_{m+i})} \tag{7.5}
$$

and

$$
\lambda^{OC}_{m+i} = \exp(\beta_0^{OC} + \beta_1^{OC} x_{m+i}). \tag{7.6}
$$

where at least one of the following holds: $\hat{\beta}_0 \neq \beta_0^{OC}, \hat{\beta}_1 \neq \beta_1^{OC}, \hat{\gamma}_0 \neq \gamma_0^{OC}$ or $\hat{\gamma}_1 \neq \gamma_1^{OC}$. Hence, the process becomes OC when dependent observations $\{Y_{m+\mathcal{T}+1}, Y_{m+\mathcal{T}+2}, \dots\}$ follow a ZIP distribution with parameters that deviate from what is expected by the established regression model, based on the observed covariate values for $\{X_{m+\mathcal{T}+1}, X_{m+\mathcal{T}+2}, \dots\}$.

It is described in Section 4.2.1 that a Shewhart chart can either have symmetric control limits as defined in (4.1), or probability control limits as defined in (4.2). Let us denote the ZIP regression-based Shewhart chart with symmetric control limits as ZIP-(r^P, L)-Shewhart, ZIP-(r^D, L)-Shewhart and ZIP-(r^Q, L)-Shewhart in case of Pearson, deviance and randomised quantile residuals respectively. Let us denote ZIP-(r^P, Q)-Shewhart, ZIP-(r^D, Q)-Shewhart and ZIP-(r^Q, Q)-Shewhart for similar control charts with probability limits.

7.1.2 The ZINB regression-based Shewhart chart

Monitoring ZINB residuals is similar as described in the previous section, since the ZIP and ZINB regression models have the same link functions. Let us assume that random variable Y_i follows a ZINB distribution. Also, let us define the Phase I data with $(y, x) = \{(y_1, x_1), \dots, (y_m, x_m)\}$. We can fit the one dimensional ZINB regression model that is defined in (6.13) to the Phase I data in order to obtain estimates for regression coefficients $\beta_0, \beta_1, \gamma_0, \gamma_1$ and τ. Here, τ represents the non-negative size parameter of the ZINB distribution. Let us denote these estimates with $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1$ and $\hat{\tau}$, respectively. For every new arriving observation (y_{m+i}, x_{m+i}) with $i = 1, 2, \dots$, predictions $\hat{\mu}_{m+i}$, $\hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$ are defined as in (7.1), (7.2) and (7.3) respectively, where the ZINB estimated regression coefficients are

applied. Notice that these predictions have the same definition in the ZIP regression model, since the ZINB and ZIP expected value are the same, as defined in (6.8) and (6.14), respectively.

ZINB Pearson, deviance and randomised quantile residuals are obtained by substituting $y_{m+i}, \hat{p}_{m+i}$ and $\hat{\lambda}_{m+i}$ in (6.16), (6.17) and (6.18), respectively. Let us use the same notation as in Section 7.1.1, i.e., $r^P = \{r^P_{m+1}, r^P_{m+2}, \dots\}$, $r^D = \{r^D_{m+1}, r^D_{m+2}, \dots\}$ and $r^Q = \{r^Q_{m+1}, r^Q_{m+2}, \dots\}$. We can monitor each of these residual types over time in a Shewhart chart. As long as the process is in control, it is assumed that observations Y_{m+i} with $i = 1, 2, \dots$ follow a ZINB distribution with parameters $\hat{p}_{m+i}$, $\hat{\lambda}_{m+i}$ and τ, i.e., $Y_{m+i} \sim \text{ZINB}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}, \tau | X_{m+i})$. Then the hypotheses of this chart are defined as

$$
\begin{aligned}
H_0: &\quad Y_{m+i} \sim \text{ZINB}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}, \hat{\tau} | X_{m+i}) \quad \text{for} \quad i = 1, 2, \dots \\
H_1: &\begin{cases} Y_{m+i} \sim \text{ZINB}(\hat{p}_{m+i}, \hat{\lambda}_{m+i}, \hat{\tau} | X_{m+i}) & \text{for} \quad i = 1, \dots, \mathcal{T} \\ Y_{m+i} \sim \text{ZINB}(p^{OC}_{m+i}, \lambda^{OC}_{m+i}, \hat{\tau} | X_{m+i}) & \text{for} \quad i = \mathcal{T}+1, \mathcal{T}+2, \dots . \end{cases}
\end{aligned} \tag{7.7}
$$

where $\hat{p}_{m+i} \neq p^{OC}_{m+i}$ and or $\hat{\lambda}_{m+i} \neq \lambda^{OC}_{m+i}$. Hence, the process becomes out of control after change point $\mathcal{T}$. For a known observation (y_{m+i}, x_{m+i}), parameters p^{OC}_{m+i} and λ^{OC}_{m+i} are defined as in (7.5) and (7.6), where at least one of the following holds: $\hat{\beta}_0 \neq \beta^{OC}_0, \hat{\beta}_1 \neq \beta^{OC}_1, \hat{\gamma}_0 \neq \gamma^{OC}_0$ or $\hat{\gamma}_1 \neq \gamma^{OC}_1$. In this notation, $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0$ and $\hat{\gamma}_1$ represent the estimated regression coefficients from the ZINB model.

Let us denote the ZINB regression-based Shewhart chart with symmetric control limits as ZINB-(r^P, L)-Shewhart, ZINB-(r^D, L)-Shewhart and ZINB-(r^Q, L)-Shewhart in case of Pearson, deviance and randomised quantile residuals, respectively. Let us denote similar control charts with probability limits as ZINB-(r^P, Q)-Shewhart, ZINB-(r^D, Q)-Shewhart and ZINB-(r^Q, Q)-Shewhart.

Throughout this section, we have defined the ZIP and ZINB regression-based Shewhart charts with both symmetric and probability control limits, and for monitoring Pearson, deviance and randomised quantile residuals. The goal of this project is to evaluate the performance of each charts, for both high and low proportions of zero-inflation in the IC data. Figure 7.1 shows a detailed overview of this solution strategy that was introduced in Section 3.4, where a distinction between symmetric and probability control limits is made.

In addition, it is described in Section 4.2.1 that Shewhart charts were originally defined for monitoring normally distributed data. However, it is explained in Section 6.1.4 that normality of Pearson, deviance and randomised quantile residuals is not guaranteed. Therefore, we evaluate the distribution of each residual type in Section 7.3. The results in this section are based on simulated data, such that the methodology for simulating zero-inflated data is described first in the following section.

7.2 Simulation of zero-inflated data depending on one covariate

We can simulate ZIP and ZINB distributed observations Y_i that depend on covariate X_i for $i = 1, 2, \dots$ directly from regression models (6.7) and (6.13), respectively. This method is also described in Mahmood (2020). In Chapter 2 it is explained that we assume $X_i \sim N(\mu_X, \sigma^2_X)$. For simulation purposes, we fix $\mu_X = 0$ and $\sigma_X = 1$. In the context of plastic pellet production, this could be achieved by normalisation of the inspected weights. Then, ZIP and ZINB data simulation for defined values of β_0, β_1, γ_0, γ_1 and τ is executed as defined in (7.8).

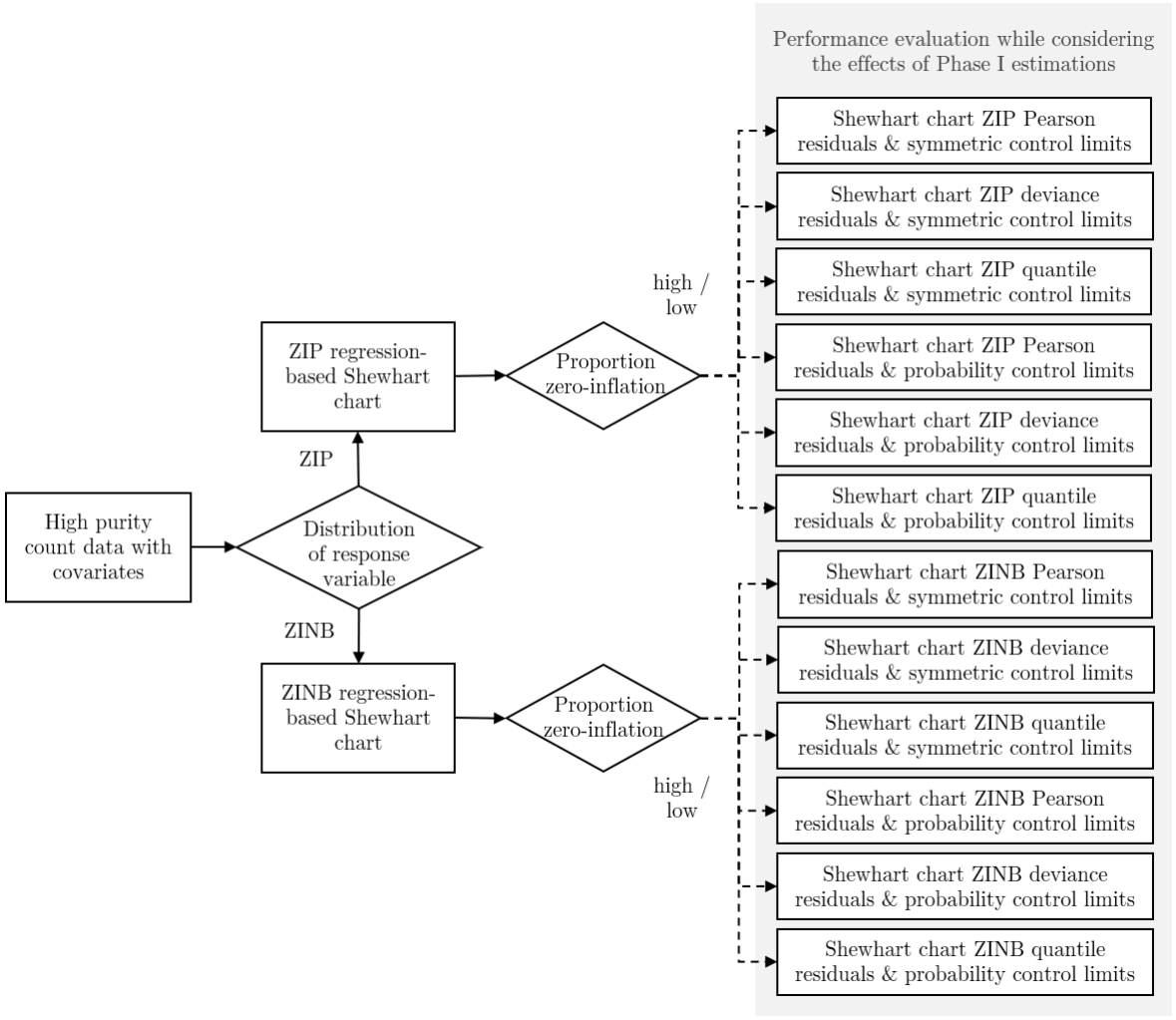

Figure 7.1: Flow chart of the detailed solution strategy.

$$X_i \sim N(0,1)\,, \quad p_i = \frac{\exp\{\gamma_0 + \gamma_1 X_i\}}{1+\exp\{\gamma_0 + \gamma_1 X_i\}}\,, \quad \lambda_i = \exp\{\beta_0 + \beta_1 X_i\}$$

$$c_i \sim \text{Bernoulli}(p_i)$$

$$\text{ZIP}: \quad \begin{cases} Y_i \sim \text{Poisson}(\lambda_i) & \text{if} \quad c_i = 0 \\ Y_i = 0 & \text{if} \quad c_i = 1 \end{cases} \tag{7.8}$$

$$\text{ZINB}: \quad \begin{cases} Y_i \sim \text{NBinom}(\lambda_i, \tau) & \text{if} \quad c_i = 0 \\ Y_i = 0 & \text{if} \quad c_i = 1 \end{cases}$$

The random Bernoulli variable c_i is simulated to indicate if Y_i equals a structural zero, or whether Y_i follows a Poisson or negative binomial distribution. The parameter p_i represents the probability that Y_i equals a structural zero. The parameter λ_i represents the expected value of the Poisson or negative binomial distribution, in case Y_i is not a structural zero. The parameter τ represents the strictly positive size parameter of the negative binomial distribution. Parameters γ_0, γ_1, β_0 and β_1 of (7.8) correspond to the regression coefficients from the ZIP and ZINB regression model in (6.7) and (6.13), respectively. Hence, the parameters γ_0, γ_1, β_0 and β_1 affect the relationship between covariates X_i and observations

Y_i.

7.2.1 Simulating four IC scenarios

The goal of this project is to compare the performance of each regression-based Shewhart chart under various IC distributions. Hence, we consider four different IC process scenarios, in which the observations Y_i follow a ZIP or ZINB distribution with either high or low proportions of zero-inflation. These four scenarios are defined by parameter values for γ_0, γ_1, β_0, β_1 and τ. Each scenario is denoted in Table 7.1. Scenario 1 represents a process where observations Y_i follow a ZIP(p_i,λ_i) distribution with a high expected proportion of zero-inflation and low expected value. Scenario 2 represents a similar process with low expected proportion of zero-inflation and high expected value. In addition, Scenario 3 represents a process where observations Y_i follow a ZINB(p_i,λ_i,τ) distribution with a high expected proportion of zero-inflation and low expected value, and Scenario 4 represents a similar process with low expected proportion of zero-inflation and high expected value.

Scenario	Distribution	μ_X	σ_X	β_0	β_1	γ_0	γ_1	τ	$E[p_i]$	$E[\lambda_i]$
1	ZIP	0	1	0.1	1.0	0.5	-1.0	-	0.60	1.82
2	ZIP	0	1	1.6	1.0	-0.6	-1.0	-	0.38	8.17
3	ZINB	0	1	0.1	1.0	0.5	-1.0	11	0.60	1.82
4	ZINB	0	1	1.6	1.0	-0.6	-1.0	3	0.38	8.17

Table 7.1: Parameter values for simulating Phase I data for the ZIP and ZINB distributions with high and low proportions of zero-inflation

The parameters p_i and λ_i are actually random variables since they depend on X_i for $i = 1, 2, \ldots$. Therefore, their expected values are provided in the last two columns of Table 7.1, as an indication of the amount of zero inflation in the IC scenario. These values are obtained according to (7.9) and (7.10), where it is used that $X_i \sim N(0, 1)$ for all $i = 1, 2, \ldots$.

$$\mathrm{E}\left[p_i\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp\left(\gamma_0 + \gamma_1 x - \frac{1}{2}x^2\right)}{1 + \exp(\gamma_0 + \gamma_1 x)} dx. \tag{7.9}$$

$$\mathrm{E}\left[\lambda_i\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\beta_0 + \beta_1 x - \frac{1}{2}x^2\right) dx. \tag{7.10}$$

Figure 7.2 shows the histograms of 1500 simulated observations Y_i in each scenario, to illustrate their distributions and respective proportions of zero-inflation. Now that we have defined the four IC scenarios, we can continue to with the construction of ZIP and ZINB regression-based Shewhart charts with Pearson, deviance and randomised quantile residuals. However, the distribution each type of residuals is analysed first in the following section.

7.3 Distribution analysis of Phase I regression residuals

In this section, we analyse the distribution of Pearson, deviance and randomised quantile residuals for each of the IC scenarios that are defined in Section 7.2.1. Let us consider IC Scenario 1, where observations Y_i follow a ZIP(p_i, λ_i) distribution with high expected proportion of zero inflation. We assume for know that all distributional parameters are known, such that the regression model for each IC scenario is fixed at the true parameters $\beta_0 = 0.1, \beta_1 = 1.0, \gamma_0 = 0.5$ and $\gamma_1 = -1.0$. Hence, we do not have a Phase I

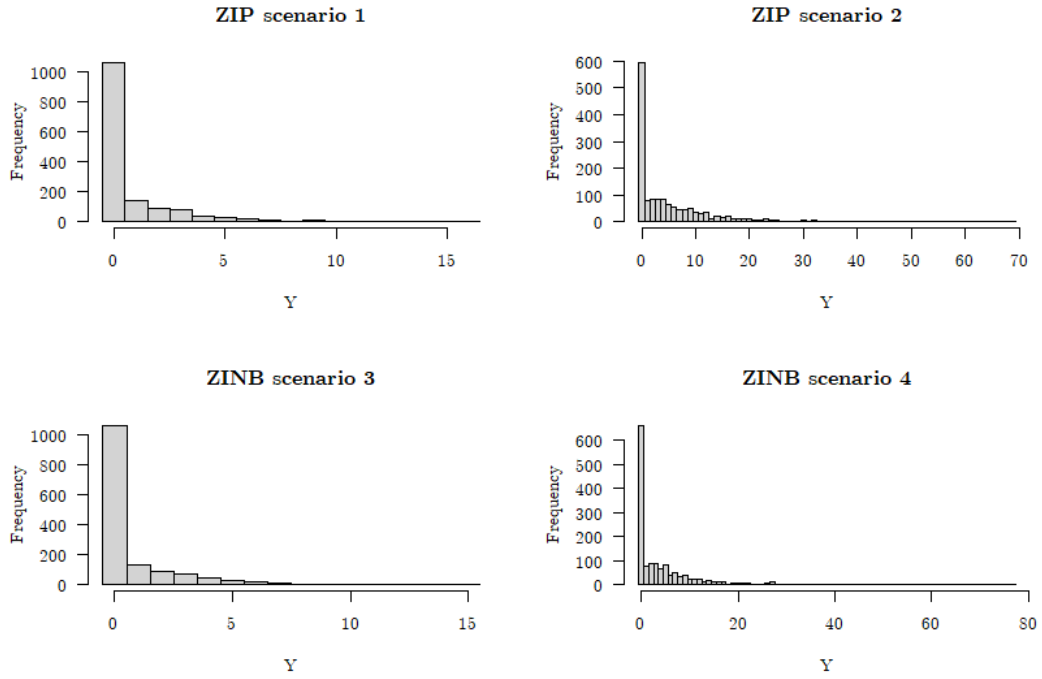

Figure 7.2: Histogram of simulated Y_i for each scenario in Table 7.1

and $m = 0$. A set of $n = 1500$ observations $(y, x) = \{(y_1, x_1), \ldots, (y_n, x_n)\}$ is simulated as described in Section 7.2. Pearson, deviance and randomised quantile residuals are obtained from the simulated data, according to definitions (6.10), (6.11) and (6.12), respectively. Distributions of each type of residuals r^P, r^D and r^Q are shown in Figure 7.3. Here, we distinguish between the residual density that originates from the Poisson distribution, and the density that originates from structural zeros in the data, i.e., the inflated zeros.

It is concluded from Figure 7.3 that randomised quantile residuals are normally distributed, as expected. However, this does not hold for Pearson and deviance residuals. It is observed from the coloured overlay in Figure 7.3 that the positive skewness of Pearson and deviance residuals is mostly caused by the residuals that correspond to the zero-inflated observations. The non-normal residual distributions are remarkable, since this is nowhere discussed in the current SPC literature on ZIP and ZINB regression-based control charts. Namely, a ZIP and ZINB regression-based Shewhart chart with Pearson residuals is studied in the research of Mahmood (2020), but the skewness of residual distribution remains unnoticed. In addition, Park et al. (2020) introduces a ZIP regression-based Shewhart chart with deviance residuals. In the research of Park et al. (2020), similar distribution of deviance residuals is observed as is shown in Figure 7.3, but the non-normality of ZIP deviance residuals is ignored and symmetric control limits are proposed nevertheless. Before we continue with constructing any type of Shewhart chart, further distribution analysis of ZIP Pearson and deviance residuals is needed.

It is described in Section 7.1 how we monitor regression residuals from a given ZIP or ZINB regression model over time in a Shewhart chart. When monitoring these regression residuals, we aim to detect contextual anomalies. Hence, we aim to identify data points (y_i, x_i) for which the predicted value $\hat{\mu}_i$ severely deviates from the observed value y_i. The Shewhart chart detects OC data points by setting control limits at the upper and lower distributional tail of the charting statistic. Hence, in order to decide whether we can use Pearson and deviance residuals in a Shewhart chart, we must discover whether severe cases of over estimation, i.e., $\hat{\mu}_i \gg y_i$ show in the upper distributional tail of the regression residuals. We must also be convinced that severe cases of under estimation, i.e., $\hat{\mu}_i \ll y_i$, show in the lower distributional

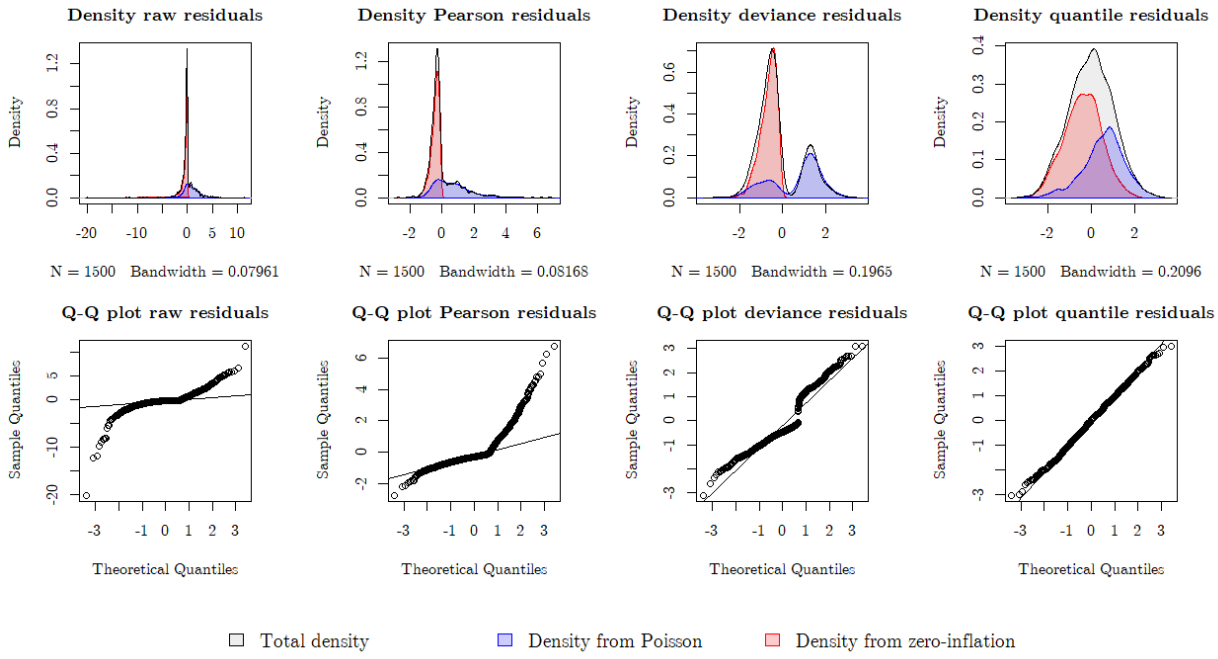

Figure 7.3: Density and Q-Q plots of ZIP regression residuals for IC Scenario 1 (ZIP), indicating whether the residuals originates from the zero-inflation (red) or the Poisson distribution (blue).

tail of the regression residuals. Therefore, a break down of ZIP Pearson and deviance residuals is shown in Figure 7.4 and 7.5, respectively.

For each observation (y_i, x_i) with $i = 1, \ldots, n$, Figure 7.4a shows the corresponding raw residual $r_i = y_i - \hat{\mu}_i$, which is plotted against the prediction $\hat{\mu}_i$. Figure 7.4b shows the Pearson residuals $r_i^P = (y_i - \hat{\mu}_i)/V(\hat{\mu}_i)$ plotted against prediction $\hat{\mu}_i$. In addition, Figure 7.4c shows the density of Pearson residuals. When comparing Figure 7.4a with Figure 7.4b, we observe that the density skewness of Pearson residuals is caused by dividing the raw residual by the variance function, in case $V(\hat{\mu}_i)$ is very low. However, Figure 7.4b shows that Pearson residuals r_i^P that correspond to severe overestimation, i.e., $\hat{\mu}_i \gg y_i$ end up in the lower tail of the density nevertheless. Similarly, the residuals that correspond to severe cases of underestimation, i.e., $\hat{\mu}_i \ll y_i$ end up in the upper tail of the density.

A similar analysis is carried out for ZIP deviance residuals in Figure 7.5. Here, Figure 7.5a shows the square root of the unit deviance, i.e., $\sqrt{d_i}$ as defined in Section 6.2.1, that corresponds to each observation (y_i, x_i), which is plotted against the predicted value $\hat{\mu}_i$ for $i = 1, \ldots, n$. It is observed that $\sqrt{d_i}$ does not approach zero when the true value of y_i is greater or equal to 1. Only for $y_i = 0$, we observe cases in which $\sqrt{d_i}$ is close to zero. Hence, the ZIP regression model seems to structurally underestimate observations that have true value $y_i > 0$. It is also described in Section 6.2.1 that the ZIP deviance residuals are defined as the signed square root of the unit deviance, i.e., $r_i^D = \text{sign}(y_i - \hat{\mu}_i)\sqrt{d_i}$. Figure 7.5b shows the deviance residuals r_i^D plotted against $\hat{\mu}_i$, where it shows that the two peeks in the distribution are clearly caused by the multiplication with the sign, and the fact that $\sqrt{d_i}$ does not approach 0 for $y_i > 0$.

However, we observe the same behaviour for deviance residuals as for Pearson residuals. Namely, the deviance residuals r_i^D that correspond to severe overestimation, i.e., $\hat{\mu}_i \gg y_i$ end up in the lower tail of the density. Also, the values r_i^D that correspond to severe cases of underestimation, i.e., $\hat{\mu}_i \ll y_i$ end up in the upper tail of the density. With these conclusions, it is confirmed that both ZIP Pearson and deviance residuals are proper measures for goodness of fit, which makes them appropriate for contextual anomaly detection. Hence, we can monitor ZIP Pearson, deviance and randomised quantile residuals in a Shewhart chart over time.

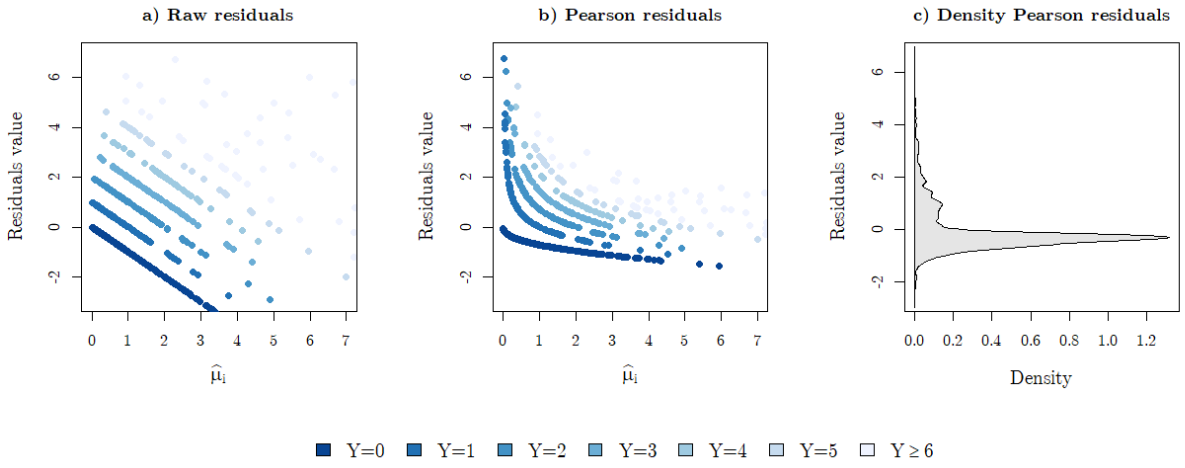

Figure 7.4: Break down of Pearson residuals in IC Scenario 1 (ZIP), with: a) Raw residuals $r_i = y_i - \hat{\mu}_i$ plotted against prediction $\hat{\mu}_i$, b) Pearson residuals r_i^P plotted against prediction $\hat{\mu}_i$ and c) a density plot of Pearson residuals.

It is explained in Section 4.2.1 that the run length of a Shewhart chart follows a geometric distribution if the charting statistic is independent and identically distributed. Therefore, autocorrelation plots of the simulated residuals are provided in Figure 7.6, to analyse whether each residual type is independently distributed.

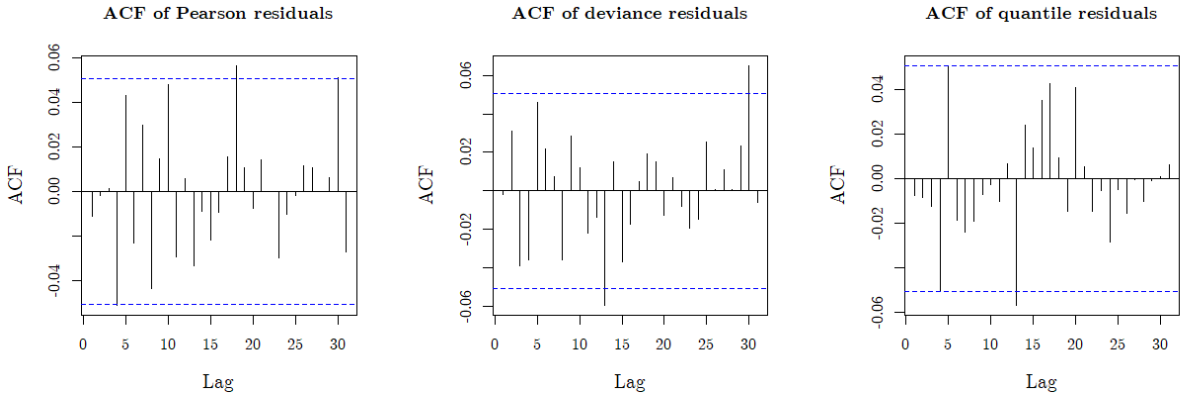

Figure 7.6: Autocorrelation plot of Pearson, deviance and randomised quantile residuals in residuals for IC Scenario 1 (ZIP).

It is observed from Figure 7.6 that all residual types show signs of autocorrelation. However, the autocorrelation function (ACF) values appear random with respect to the time lags. Notice that these ACF results are obtained for 1500 simulated residuals, which causes variation in the ACF results. Let us from now on assume that the ZIP randomised quantile residuals follow an independent and identical standard normal distribution. This assumption is not necessarily true, as is slightly indicated by Figure 7.6, but it simplifies construction of the (r^Q, L)- and (r^Q, Q)-Shewhart charts significantly. This is further discussed in Section 7.5. In addition, it is decided to make no assumptions regarding the distribution of ZIP Pearson and deviance residuals, since Figures 7.3 and 7.6 show no convincing distribution or signs of independence. Hence, we will approach the construction of each (r^P, L)-, (r^D, L)-, (r^P, Q)- and (r^D, Q)-Shewhart chart numerically. This is also further discussed in Section 7.5.

A similar analysis is conducted for all other IC scenarios. The residual distribution and ACF plots for each of these scenarios are attached in Appendix B.2. Similar conclusions are obtained from each analysis,

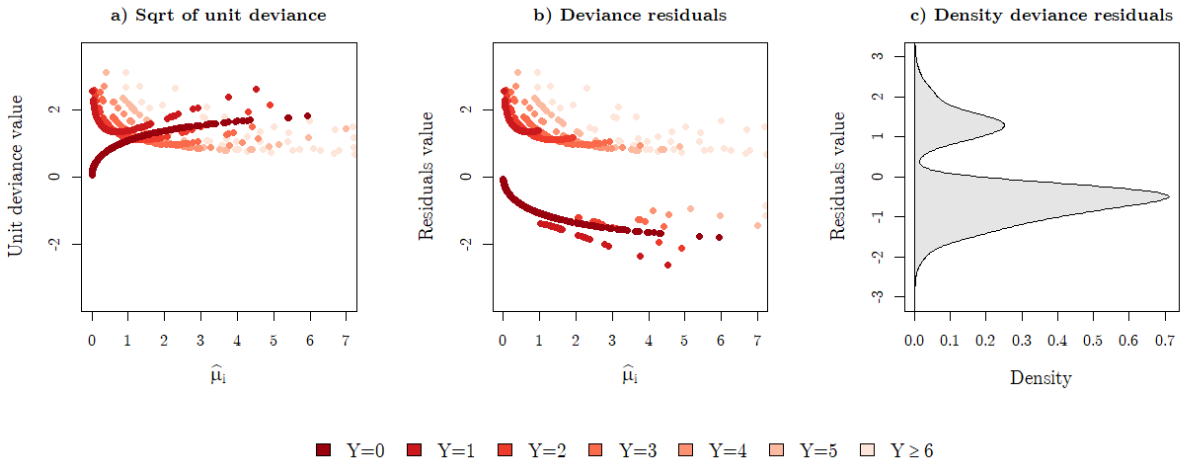

Figure 7.5: Break down of deviance residuals in IC Scenario 1 (ZIP), with: a) the squared unit deviance $\sqrt{d_i}$ plotted against prediction $\hat{\mu}_i$, b) deviance residuals r_i^D plotted against prediction $\hat{\mu}_i$ and c) a density plot of deviance residuals.

as are discussed in this section. Hence, we assume randomised quantile residuals from a ZIP or ZINB regression model follow a independent and identical standard normal distribution, and no assumptions are made regarding the distributions of Pearson and deviance residuals throughout this project. In the following section, it is discussed how each ZIP and ZINB regression-based Shewhart charts is constructed, according to two distinct strategies for performance evaluation.

7.4 Two strategies for performance evaluation

The goal of this project is to provide Dow with insights in the performance of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart for both ZIP and ZINB distributed data. From the literature review in Section 3.2 it is known that two studies have been published on the performance of the ZIP and ZINB regression-based Shewhart chart. However, Park et al. (2020) and Mahmood (2020) only provide us with the baseline performance of the ZIP and ZINB regression-based Shewhart charts. These results reflect the best possible performance of each chart, since results are obtained under the assumption of perfect model fit. However, it is more interesting to understand what the true performance of each control chart is, while taking into account the effects estimating the Phase I regression coefficients. This because having a Phase I is inevitable when monitoring predictive residuals for a real life production process. When taking into account the effect of having Phase I estimates, it is possible that the true ARL_0 will is lower than the intended ARL_0 (see e.g. Albers and Kallenberg (2004) and Shu et al. (2005)).

Therefore, two strategies of performance analysis are considered in this thesis. The first strategy is to establish the baseline performance of each chart, whereas it is assumed that all parameters are known. This corresponds to a performance evaluation in which we ignore the effects of Phase I estimation. In the second performance evaluation strategy, we estimate the effect of the Phase I estimation. Hence, results from this analysis reflect the true performance of each Shewhart chart, while taking into account that we have a Phase I. Both strategies are discussed in the following sections.

7.4.1 Performance evaluation while ignoring the Phase I effects

The fist strategy is the most simplistic, in which we assume all data properties are known. Therefore, we do not consider a Phase I, i.e. $m = 0$. In case observations are ZIP distributed, then the ZIP regression

model is applied as defined in (6.7), with fixed regression coefficients $\beta_0, \beta_1, \gamma_0$ and γ_1. In case observations are ZINB distributed, then the ZINB regression model is applied as defined in (6.13), with fixed regression coefficients $\beta_0, \beta_1, \gamma_0$ and γ_1 and additional size parameter τ. All parameter values are defined for each IC scenario in Table 7.1. Covariate X_i is assumed to follow a standard normal distribution, i.e., $\mu_X = 0$ and $\sigma_X = 1$. The steps of performance evaluation in a setting where all data properties are assumed to be known, are defined as follows:

1. **Construct the Shewhart chart:** We construct the Shewhart chart to achieve an ARL_0 of 200, by determining L or Q_1 and Q_2. The procedure for obtaining symmetric and probability control limits is described in Section 7.5. Known parameters $\mu_X, \sigma_X, \beta_0, \beta_1, \gamma_0, \gamma_1$ and τ are applied in case L or Q_1 and Q_2 are numerically solved by simulation.

2. **Evaluate performance by simulating ARL_1:** The performance of each chart is evaluated upon the ARL_1, that is obtained by simulation. The simulation of OC data is described in Section 7.6.1, for which parameters $\mu_X, \sigma_X, \beta_0, \beta_1, \gamma_0, \gamma_1$ and τ are used as the baseline from which OC parameters β_0^{OC}, β_1^{OC}, γ_0^{OC} and γ_1^{OC} are determined. The steps for obtaining the ARL_1 are described in Section 7.6.2.

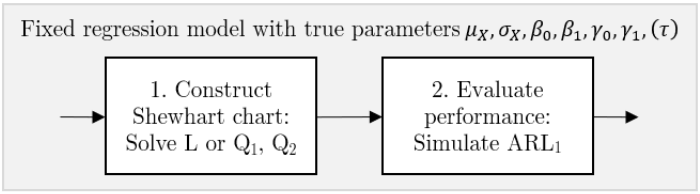

Figure 7.7: Graphical representation of performance evaluation strategy, when ignoring the Phase I effects.

A graphical representation of these two steps is shown in Figure 7.7. It is assumed that all IC parameters are known in this performance evaluation strategy, such that the best possible performance of each chart is obtained. Hence, we call this the baseline performance. The alternative performance evaluation strategy takes into account the effects of Phase I estimation, which is described in the following section.

7.4.2 Performance evaluation while estimating the Phase I effects

In the previous section it is assumed that all IC parameters are known. This is however never the case in practice. Instead, IC parameters are estimated from a stable Phase I period. The effect of Phase I estimation is taken into account in this second performance evaluation strategy. Shu et al. (2005) follows a similar performance evaluation strategy, of which the consecutive steps are defined as:

1. **Phase I data simulation:** We simulate a Phase I data set with m observations $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$. Simulation of ZIP or ZINB data is described in Section 7.2, for which the true parameters μ_X, σ_X, β_0, β_1, γ_0, γ_1 and τ are used. True parameter values are defined per IC scenario in Table 7.1.

2. **Fit the Phase I regression model:** The ZIP or ZINB regression model is fitted to the Phase I data, as defined in (6.7) and (6.13), respectively. Estimated parameters are denoted as $\bar{x}$, s_x, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\tau}$.

3. **Construct the Shewhart chart:** We construct the Shewhart chart to achieve an ARL_0 of 200, by determining L or Q_1 and Q_2. The procedure for obtaining symmetric and probability control limits is described in Section 7.5. Estimated parameters $\bar{x}$, s_x, $\hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\tau}$ are applied in case

L or Q_1 and Q_2 are numerically solved by simulation. Hence, each Shewhart chart is constructed according to the estimated IC distribution parameters.

4. **Evaluate performance by simulating the true ARL_0 and ARL_1:** The performance of each chart is evaluated upon the true ARL_0 and ARL_1, which are both obtained by simulation. The simulation of OC data is described in Section 7.6.1, for which true parameters $\mu_X, \sigma_X, \beta_0, \beta_1, \gamma_0, \gamma_1$ and τ are used as the baseline from which OC parameters β_0^{OC}, β_1^{OC}, γ_0^{OC} and γ_1^{OC} are determined. The steps for obtaining the ARL_1 are described in Section 7.6.2. Hence, the true performance of each control chart is estimated according to the true IC distribution parameters.

5. **Repeat:** Steps 1-4 are repeated 100 times.

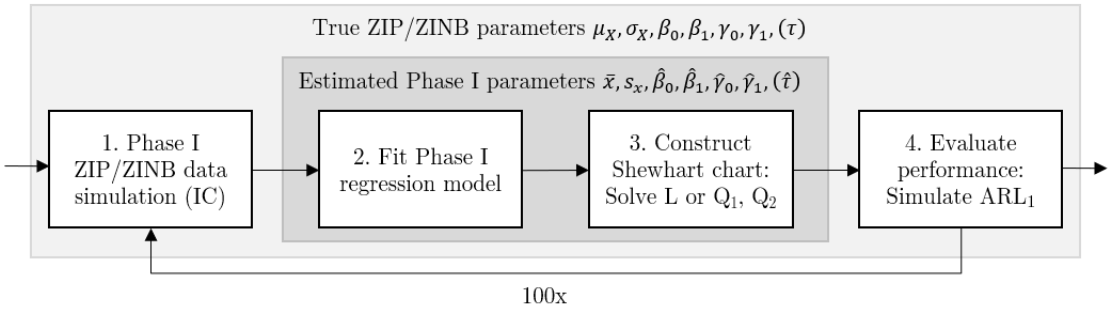

Figure 7.8: Graphical representation of performance evaluation strategy, when estimating the effect of Phase I.

Steps 1-4 are repeated to eliminate the variation that is inherited in the Phase I data. The ARL results from step 4 are accumulated and averaged over all 100 iterations, to obtain the final results. The $SDRL$ results are obtained by computing the standard deviation of all computed run lengths, as well as by taking the square root of the pooled variance of the 100 iterations. Interpretation of both $SDRL$ results is further discussed in Section 8.2. A graphical representation of these two steps is shown in Figure 7.8. The ARL_1 results of this strategy reflect the true performance of the ZIP and ZINB regression-based Shewhart charts, since the effect of Phase I estimations is taken into account. Now that we have defined two distinct methods for performance evaluation, we continue with the methodology for obtaining control limits in the following section.

7.5 Constructing the regression-based Shewhart chart

It is described in Section 7.1 that ZIP and ZINB regression residuals are monitored over time in a Shewhart chart with either symmetric or probability control limits. The conclusion from Section 7.3 states that both Pearson and deviance residuals follow a unknown distribution, in case of ZIP and ZINB regression. No assumptions are made regarding independence of Pearson and deviance residuals as well, such that control limits are solved numerically. Randomised quantile residuals, on the other hand, are assumed to follow an independent and identical standard normal distribution. Under this assumption, the run length of the (r^Q, L)- and (r^Q, Q)-Shewhart charts is geometrically distributed. Hence, the (r^Q, L)-Shewhart chart will achieve an ARL_0 of 200 for $L = 2.81$, since $1-P(LCL < r^Q < UCL) = 2(1-\Phi(2.81)) \approx 1/200$. With similar reasoning, the (r^Q, Q)-Shewhart chart will also achieve an ARL_0 of 200 for $Q_1 = 2.81$ and $Q_2 = -2.81$. A regression-based control chart with quantile residuals is therefore avoiding the need for numerically solved control limits. This can be a significant advantage in practice. The procedure for solving charting constants L, Q_1 and Q_2 numerically for the Shewhart charts with Pearson and deviance residuals is discussed in Section 7.5.1.

7.5.1 Solving charting constants L, Q_1 and Q_2 numerically

Let us consider again the Phase I data set $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$ where random variable Y_i follows a ZIP or ZINB distribution. Phase I estimated parameters are denoted with $\bar{x}$, s_x, $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1$, and $\hat{\tau}$ in case Y_i follows a ZINB distribution. The Phase I ZIP and ZINB regression models are defined in (6.7) and (6.13), respectively. The charting constant L is solved for the (r^P, L)- and (r^D, L)-Shewhart chart to obtain an ARL_0 of 200. Similarly, charting constants Q_1 and Q_2 are solved for the (r^P, Q)- and (r^D, Q)-Shewhart chart to obtain an ARL_0 of 200. Both solving procedures are the same for Pearson and deviance residuals, such that the general notation r_i is applied in this section to denote the obtained residual at time i. The consecutive steps are defined as follows:

1. An IC Phase II data set is generated with N runs of n ZIP or ZINB distributed observations. Let us denote these observations with $(y_{\ell,m+i}, x_{\ell,m+i})$, where m is the size of the Phase I data, $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$. Parameters $\bar{x}, s_x, \hat{\beta}_0$, $\hat{\beta}_1$, $\hat{\gamma}_0$, $\hat{\gamma}_1$ and $\hat{\tau}$ are applied in the simulation, which is defined in (7.8).

2. The established Phase I ZIP or ZINB regression model is applied to obtain the regression residuals from each observation $(y_{\ell,m+i}, x_{\ell,m+i})$. Let us denote each run of residuals with $R_\ell = \{r_{\ell,m+1}, \ldots, r_{\ell,m+n}\}$ with $\ell = 1, \ldots, N$.

3. **In case of symmetric Shewhart chart, obtain L:** A Shewhart chart with symmetric control limits is constructed as defined in (4.1), with control limits $\mu_r \pm L\sigma_r$. Here, μ_r and σ_r denote the mean and standard of the simulated residuals from Step 2. For an arbitrary value of L, the run length of each run of residuals R_ℓ is determined and denoted with RL_ℓ, for $\ell = 1, \ldots, N$. The ARL_0 is determined as the average of all computed run lengths, i.e. $ARL_0 = (RL_1 + \cdots + RL_N)/N$. The charting constant L is obtained by solving the control limit equation in (4.1) to achieve $ARL_0 = 200$.

3. **In case of Shewhart chart with probability limits, obtain Q_1 and Q_2:** A Shewhart chart with probability control limits is constructed as defined in (4.2). For $\alpha = 1/200$, the limits Q_1 and Q_2 are obtained ensuring that $100(1 - \alpha/2)$ percent of the simulated residuals $r_{\ell,m+i}$ with $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$ is larger that Q_1. Similarly, Q_2 is obtained to achieve that $100(\alpha/2)$ percent of the simulated residuals is smaller that Q_2.

Simulation size parameters are set to $n = 3000$ and $N = 10,000$ when evaluating baseline performance. There is no Phase I in this case since all parameters are assumed to be known, such that $m = 0$ and $\bar{x} = \mu_X$, $s_x = \sigma_X$, $\hat{\beta}_0 = \beta_0, \hat{\beta}_1 = \beta_1, \hat{\gamma}_0 = \gamma_0, \hat{\gamma}_1 = \gamma_1$, and $\hat{\tau} = \tau$. Values for L that are obtained for baseline performance are provided in Table 7.2. Values for Q_1 and Q_2 that are obtained for baseline performance are provided in Table 7.3.

IC	(r^P, L)-Shewhart			(r^D, L)-Shewhart			(r^Q, L)-Shewhart		
scenario	μ_{r^P}	σ_{r^P}	L	μ_{r^D}	σ_{r^D}	L	μ_{r^Q}	σ_{r^Q}	L
1 (ZIP)	0.0001	1.0003	4.6233	-0.1799	1.0192	2.6843	0.0000	1.0000	2.8100
2 (ZIP)	0.0002	1.0002	3.3353	-0.1338	1.2740	2.1642	0.0000	1.0000	2.8100
3 (ZINB)	0.0001	1.0004	4.7200	-0.1993	1.0018	2.7260	0.0000	1.0000	2.8100
4 (ZINB)	0.0003	1.0003	3.9734	-0.3100	1.2039	2.2902	0.0000	1.0000	2.8100

Table 7.2: Obtained charting constants L for the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart, and for each IC scenario from Table 7.1, in case of baseline performance evaluation.

Simulation size parameters are set to $m = 1500$, $n = 3000$ and $N = 200$ when evaluating performance while taking into account the effects of Phase I estimation, since 100 replications are executed in this

IC scenario	Obtained probability limits (Q_1, Q_2) for (r^P, Q)-Shewhart	(r^D, Q)-Shewhart	(r^Q, Q)-Shewhart
1 (ZIP)	(5.6328, -1.7013)	(2.7749, -2.0966)	(2.8100, -2.8100)
2 (ZIP)	(3.7637, -2.7368)	(2.7769, -2.6277)	(2.8100, -2.8100)
3 (ZINB)	(5.7569, -1.4850)	(2.7486, -2.0537)	(2.8100, -2.8100)
4 (ZINB)	(4.6018, -1.4190)	(2.6673, -2.4503)	(2.8100, -2.8100)

Table 7.3: Obtained probability limits Q_1 and Q_2 for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart, and for each IC scenario from Table 7.1, in case of baseline performance evaluation.

case. The obtained for L, Q_1 and Q_2 values are different in each of the 100 results, since they are affected by the simulated Phase I data of each computation. The density of value L for the (r^P, L)- and (r^D, L)-Shewhart chart are shown in Figure B.17 for each IC scenario. Similarly, densities for Q_1 and Q_2 of the (r^P, Q)-Shewhart chart with Pearson residuals are shown in Figure B.18, and densities for Q_1 and Q_2 of the (r^D, Q)-Shewhart chart with deviance residuals are shown in Figure B.19, for each IC scenario. Argumentation for the choices of m, n and N is discussed in the following section.

7.5.2 Size of Phase I and simulation setup

The design of ARL simulations is based on a trade off between precision and computation time, as is also explained in Section 5.2.2 for the ZIP-EWMA chart. Namely, a design choice was made to generate all observations at once as a $N \times n$ matrix. It is stated in Schaffer and Kim (2007), that $\geq 6,500$ replications is enough to obtain reliable ARL_0 estimations in a Shewhart charts, and the required number of replications for reliable ARL_1 results decreases as the OC distributional shift gets larger. Since we are dealing with unknown run length distributions in the control charts for Pearson and deviance residuals, $SDRL_0$ results of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart are analysed for various N values. Figure 7.9 shows that for the $SDRL_0$ results stabilise in all charts after $N = 10,000$. It is therefore decided to execute at least 10,000 replications for each ARL computation. Hence, $N = 10,000$ in case of baseline performance evaluation. In case of performance evaluation while taking into account the effect of Phase I estimation, it is chosen to fix $N = 200$. This leads to a total of $20,000$ replications since this performance evaluation already includes a loop of 100 Phase I replications.

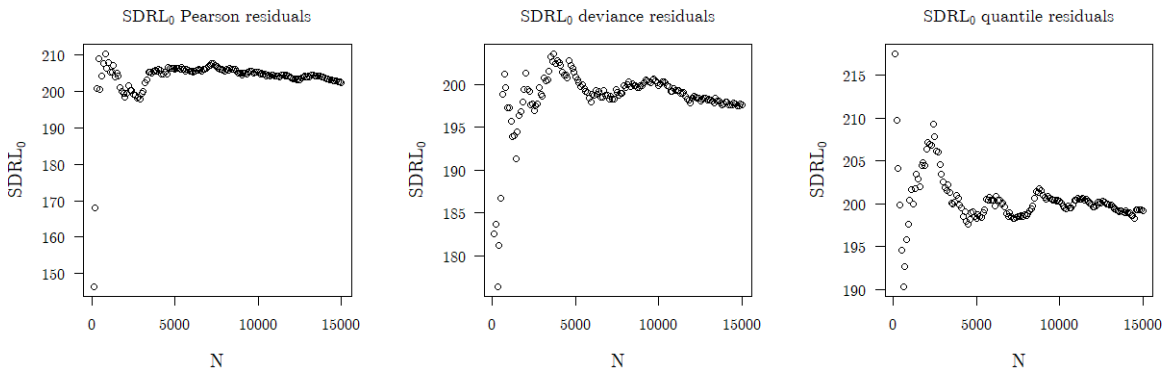

Figure 7.9: $SDRL_0$ of ZIP regression residuals for IC Scenario I, as a function of simulation size parameter N, with $N = 100, 200, \ldots, 15,000$.

In addition, we follow a similar approach for determining the total number of observations per run n, as described in Section 5.2.2. Hence, n is chosen to ensure less than 0.01 percent of all runs has no OC signal in the entire run. Simulations are executed for $n = 2000, 3000, 4000$, and for every ZIP and

ZINB regression-based Shewhart chart with Pearson, deviance and randomised quantile residuals. The results of these simulations are shown in Tables B.2 and B.3 for the Shewhart charts with symmetric and probability control limits, respectively. Results show that $n = 3000$ is large enough to ensure less than 0.01 percent of all runs leads to no OC signal.

Finally, we consider the choice of parameter m which denotes the size of the Phase I data. IC Phase I data is not always available in large amounts such that it is ideal to choose m small in practice. However, m should be large enough to ensure a good fit the ZIP and ZINB regression models. Lambert (1992) states that m must be at least 100 to obtain a reliable ZIP regression model. However, it is not stated how the required Phase I size m relates to the proportion of zero-inflation in the IC data. In addition, a required size m for the ZINB regression model is not provided by Heilbron (1994). Therefore, we determine the goodness of fit of the ZIP and ZINB regression model by simulation. Hence, for each IC scenario and for $m = 100, \ldots, 3000$, the following steps are executed:

1. A Phase I data set of size m, with ZIP or ZINB distributed observations $\{(y_1, x_1), \ldots, (y_m, x_m)\}$.

2. The ZIP or ZINB regression model is fitted to the Phase I data to estimate regression coefficients. These regression models are defined in (6.7) and (6.13), respectively.

3. An IC Phase II data set $\{(y_{m+1}, x_{m+1}), \ldots, (y_{m+n}, x_{m+n})\}$ is simulated with $n = 3000$. Predictions $\{\hat{\mu}_{m+1}, \ldots, \hat{\mu}_{m+n}\}$ are obtained from the established regression model, as described in Section 7.1. These predictions are used to obtain the unit deviance $d(y_{m+i}, \hat{\mu}_{m+i})$ for $i = 1, \ldots n$, as defined in Sections 6.2.1 and 6.2.2 for ZIP and ZINB regression, respectively. Finally, the total deviance $D(y, \hat{\mu}) = \sum_{i=1}^{n} d(y_{m+i}, \hat{\mu}_{m+i})$ is obtained as the overall goodness of fit measure.

4. Steps 1-3 are repeated 200 times to correct for the variance that is inherited in the Phase I data of Step 1. The obtained total deviance results from Step 3 are averaged over all replications.

The resulting average total deviance for each value of m with $m = 100, \ldots, 3000$ is shown in Figure 7.10, for each IC scenario. It is observed in each scenario that the total deviance stabilises after approximately $m = 1500$. The size of each ZIP and ZINB Phase I data set is therefore fixed at 1500 in this project. This concludes the construction of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart. In the following section, it is described how the OC performance of each chart is evaluated.

7.6 Performance analysis of regression-based Shewhart charts

In the previous section it is described how we construct the ZIP and ZINB regression-based Shewhart charts to obtain an ARL_0 of 200. In this section, we compare the performance of each chart by simulation of the ARL_1. The method for simulating OC ZIP and ZINB data is discussed in Section 7.6.1, after which the step-by-step approach of ARL_1 simulations are explained in Section 7.6.2.

7.6.1 Out-of-control data simulation

It is described in Section 7.1 that we assume the observations Y_i with $i = 1, 2, \ldots$ follow a ZIP(p_i,λ_i) or ZINB(p_i,λ_i,τ) distribution, where the effect of covariate X_i is modelled through parameters p_i and λ_i as defined in (7.2) and (7.3), respectively. Regression parameters β_0, β_1, γ_0 and γ_1 are either estimated from the IC Phase I data, or assumed to be known. Table 7.1 provides the true parameter values for each IC scenario. It is described in the control chart hypotheses (7.4) and (7.7) that the IC process becomes OC after time $\mathcal{T}$ when at least one of the following holds: $\beta_0 \neq \beta_0^{OC}, \beta_1 \neq \beta_1^{OC}, \gamma_0 \neq \gamma_0^{OC}$ or $\gamma_1 \neq \gamma_1^{OC}$. Let us assume that OC scenarios are not affecting the distribution of covariate X_i or the relation between X_i and observations Y_i. Hence, we assume $\mu_X^{OC} = \mu_X = 0.0$, $\sigma_X^{OC} = \sigma_X = 1.0$, $\gamma_1^{OC} = \gamma_1$ and $\beta_1^{OC} = \beta_1$.

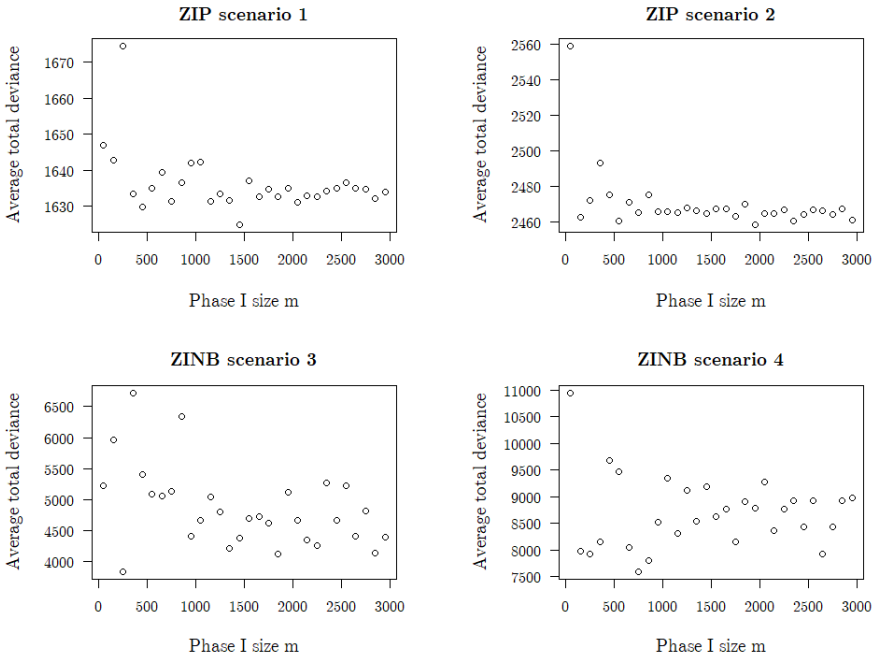

Figure 7.10: Average total deviance of the ZIP and ZINB regression model, for Phase I size $m = 100, \ldots, 3000$, for each IC scenario.

In the context of plastic pellet production at Dow, the total number of defective pellets is denoted with Y_i and the inspected weight with X_i at time $i = 1, 2, \ldots$. Overall process performance deteriorates when more defects are detected for the same inspected weight. Process performance improves when less defects are detected for the same inspected weight. In both cases, it is desirable to have an OC alarm as fast as possible, in order to start an investigation for identifying the root cause to improve future decision making. Hence, both OC scenarios with deteriorated and improved process performance are considered while evaluating the performance of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart charts.

Furthermore, we aim to simulate OC data that is proportional to the IC scenario. It is also more intuitive to define an OC scenario in terms of $E[p^{OC}]$ and $E[\lambda^{OC}]$, than in terms of β_0^{OC} and γ_0^{OC}. This because $E[p^{OC}]$ represents the expected proportion of structural zeros in the OC data, while $E[\lambda^{OC}]$ denotes the expected amount of detected defects, in case the observation Y_i is not a structural zero. The expected values $E[p^{OC}]$ and $E[\lambda^{OC}]$ are defined as

$$\mathrm{E}\left[p^{OC}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \frac{\exp\left(\gamma_0^{OC} + \gamma_1 x - \frac{1}{2}x^2\right)}{1 + \exp(\gamma_0^{OC} + \gamma_1 x)} dx \tag{7.11}$$

and

$$\mathrm{E}\left[\lambda^{OC}\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp\left(\beta_0^{OC} + \beta_1 x - \frac{1}{2}x^2\right) dx \tag{7.12}$$

The following four types of OC scenarios are simulated:

- Worse process performance due to decreased $E[p^{OC}]$, i.e. $E[p^{OC}] < E[p^{IC}]$. Equation (7.11) is solved to obtain values for γ_0^{OC} that achieve $E[p^{OC}] = \alpha_1 \cdot E[p^{IC}]$ for $\alpha_1 = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.
- Worse process performance due to increased $E[\lambda^{OC}]$, i.e. $E[\lambda^{IC}] > E[\lambda^{IC}]$. Equation (7.12) is solved to obtain values for β_0^{OC} that achieve $E[p^{OC}] = \alpha_1 \cdot E[p^{IC}]$ for $\alpha_2 = 1.3, 1.6, 1.9, 2.2, 2.5, 2.8, 3.1$.
- Improved process performance due to increased $E[p^{OC}]$, i.e. $E[p^{OC}] > E[p^{IC}]$. Equation (7.11) is

solved to obtain values for γ_0^{OC}that achieve $E[p^{OC}] = \alpha_1 \cdot E[p^{IC}]$ for $\alpha_1 = 1.05, 1.10, 1.15, 1.20, 1.25, 1.30, 1.35$.

- Improved process performance due to decreased $E[\lambda^{OC}]$, i.e. $E[\lambda^{IC}] < E[\lambda^{IC}]$. Equation (7.12) is solved to obtain values for β_0^{OC}that achieve $E[p^{OC}] = \alpha_1 \cdot E[p^{IC}]$ for $\alpha_2 = 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9$.

The ZIP Scenario 1 and ZINB Scenario 3 have $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$. ZIP Scenario 2 and ZINB Scenario 4 have $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.16$. The obtained parameters $\beta_0^{OC}$and $\gamma_0^{OC}$for each OC Scenario are provided in Table 7.4. The OC data is simulated as defined in (7.2) with parameters $\{\beta_0^{OC}, \beta_1, \gamma_0, \gamma_1, \tau, \mu_X, \sigma_X\}$, or parameters $\{\beta_0, \beta_1, \gamma_0^{OC}, \gamma_1, \tau, \mu_X, \sigma_X\}$. The ARL_1 performance of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart is simulated for every OC scenario in Table 7.4. The methodology for obtaining ARL_1 is described int the next section.

7.6.2 Performance comparison of regression-based Shewhart charts

In this section it is described how ARL_1 simulations are carried out, to compare the OC performance of each control chart with defined UCL and LCL. Let us denote the Phase I data set again with $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$ where random variable Y_i follows a ZIP or ZINB distribution. Phase I estimated parameters are denoted with $\bar{x}$, s_x, $\hat{\beta}_0, \hat{\beta}_1, \hat{\gamma}_0, \hat{\gamma}_1$, and $\hat{\tau}$ in case Y_i follows a ZINB distribution. The Phase I ZIP and ZINB regression model is defined as in (6.7) and (6.13), respectively. The methodology for simulating the ARL_1 is the same for Pearson and deviance residuals, such that the general notation r_i is again applied in this section to denote the obtained residual at time i. The consecutive steps to obtain the ARL_1 are defined as:

1. An OC Phase II data set is generated with N runs of n ZIP or ZINB distributed observations. Let us denote these observations with $(y_{\ell,m+i}, x_{\ell,m+i})$, where m is the size of the Phase I data, $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$. The OC parameters from Table 7.4 are applied in the simulation procedure that is defined in (7.8). We assume that the change point $\mathcal{T} = m$, where m denotes the size of the Phase I data set.

2. The Phase I ZIP or ZINB regression model is applied to obtain the regression residuals from each observation $(y_{\ell,m+i}, x_{\ell,m+i})$. Let us denote each run of residuals with $R_\ell = \{r_{\ell,m+1}, \ldots, r_{\ell,m+n}\}$ with $\ell = 1, \ldots, N$.

3. The run length of each run of residuals R_ℓ is determined according to the UCL and LCL of the defined control chart. These run lengths are denoted with RL_ℓ, for $\ell = 1, \ldots, N$. The ARL_1 is determined as the average of all computed run lengths, i.e. $ARL_1 = (RL_1 + \cdots + RL_N)/N$. The standard deviation of the OC run length, i.e., $SDRL_1$ is additionally obtained.

These ARL_1 and $SDRL_1$ simulations are executed for the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart, with simulation size $N = 10,000$ and $n = 3000$, and for each IC ZIP and ZINB scenario. In addition, ARL_1 values are simulated for all OC scenarios denoted in Table 7.4. The OC performance results are provided and discussed in the following chapter.

7.7 Summary

In this chapter we consider monitoring methods for zero-inflated count data that is affected by one covariate. The ZIP and ZINB regression-based Shewhart charts are defined in Section 7.1, where the hypotheses of each chart are also introduced. Simulation of zero-inflated data that depends on one covariate is discussed in Section 7.2, after which the distribution of ZIP and ZINB Pearson, deviance and randomised quantile residuals is analysed in Section 7.3. Based on this analysis it is assumed that ZIP and ZINB randomised quantile residuals follow a independent and identical standard normal distribution.

Since Figures 7.3 and 7.6 show no convincing distribution or signs of independence, it is decided to make no assumptions regarding distribution of Pearson and deviance residuals from the ZIP and ZINB regression models. It is described in Section 7.4 that two distinct strategies are consider for the OC performance evaluation of the regression-based Shewhart charts. First we establish the baseline performance of each chart, after which the true performance is estimated while taking into account the effect of having a Phase I. Construction of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart is discussed afterwards in Section 7.5. Finally, the methodology for OC performance evaluation is described in Section 7.6. The ARL_1 and $SDRL_1$ results of each regression-based Shewhart chart are presented for all IC scenarios in the following chapter.

Phase I scenario 1 (ZIP) and 3 (ZINB)									
IC parameters: $\beta_0 = 0.1, \beta_1 = 1.0, \gamma_0 = 0.5, \gamma_1 = -1.0, \tau = 11$									
$E[p^{IC}]$	α_1	$\alpha_1 \cdot E[p^{IC}]$	γ_0^{OC}	$E[p^{OC}]$	$E[\lambda^{IC}]$	α_2	$\alpha_2 \cdot E[\lambda^{IC}]$	β_0^{OC}	$E[\lambda^{OC}]$
0.6020	1.00	0.6020	0.5000	0.6020	1.8221	1.00	1.8221	0.1000	1.8221
	0.90	0.5418	0.2028	0.5418		0.90	1.6399	-0.0054	1.6399
	0.80	0.4816	-0.0890	0.4816		0.80	1.4577	-0.1231	1.4577
	0.70	0.4214	-0.3831	0.4214		0.70	1.2755	-0.2567	1.2755
	0.60	0.3612	-0.6877	0.3612		0.60	1.0933	-0.4108	1.0933
	0.50	0.3010	-1.0127	0.3010		0.50	0.9111	-0.5931	0.9111
	0.40	0.2408	-1.3731	0.2408		0.40	0.7288	-0.8163	0.7289
	0.30	0.1806	-1.7946	0.1806		0.30	0.5466	-1.1040	0.5466
	1.00	0.6020	0.5000	0.6020		1.00	1.8221	0.1000	1.8221
	1.05	0.6321	0.6532	0.6321		1.30	2.3688	0.3624	2.3688
	1.10	0.6622	0.8111	0.6622		1.60	2.9154	0.5700	2.9154
	1.15	0.6923	0.9753	0.6923		1.90	3.4620	0.7419	3.4620
	1.20	0.7224	1.1478	0.7224		2.20	4.0087	0.8885	4.0086
	1.25	0.7525	1.3308	0.7525		2.50	4.5553	1.0163	4.5553
	1.30	0.7826	1.5279	0.7826		2.80	5.1019	1.1296	5.1019
	1.35	0.8127	1.7437	0.8127		3.10	5.6486	1.2314	5.6486
Phase I scenario 2 (ZIP) and 4 (ZINB)									
IC parameters: $\beta_0 = 1.6, \beta_1 = 1.0, \gamma_0 = -0.6, \gamma_1 = -1.0, \tau = 3$									
$E[p^{IC}]$	α_1	$\alpha_1 \cdot E[p^{IC}]$	γ_0^{OC}	$E[p^{OC}]$	$E[\lambda^{IC}]$	α_2	$\alpha_2 \cdot E[\lambda^{IC}]$	β_0^{OC}	$E[\lambda^{OC}]$
0.3782	1.00	0.3782	-0.6000	0.3782	8.1662	1.00	8.1662	1.6000	8.1662
	0.90	0.3404	-0.7971	0.3404		0.90	7.3496	1.4946	7.3494
	0.80	0.3026	-1.0038	0.3026		0.80	6.5329	1.3769	6.5329
	0.70	0.2648	-1.2241	0.2648		0.70	5.7163	1.2433	5.7163
	0.60	0.2269	-1.4634	0.2269		0.60	4.8997	1.0892	4.8997
	0.50	0.1891	-1.7298	0.1891		0.50	4.0831	0.9068	4.0831
	0.40	0.1513	-2.0364	0.1513		0.40	3.2665	0.6837	3.2665
	0.30	0.1135	-2.4078	0.1135		0.30	2.4499	0.3960	2.4498
	1.00	0.3782	-0.6000	0.3782		1.00	8.1662	1.6000	8.1662
	1.05	0.3971	-0.5042	0.3971		1.30	10.6160	1.8624	10.6160
	1.10	0.4161	-0.4097	0.4161		1.60	13.0659	2.0700	13.0657
	1.15	0.4350	-0.3163	0.4350		1.90	15.5157	2.2419	15.5157
	1.20	0.4539	-0.2238	0.4539		2.20	17.9656	2.3885	17.9657
	1.25	0.4728	-0.1318	0.4728		2.50	20.4154	2.5163	20.4155
	1.30	0.4917	-0.0402	0.4917		2.80	22.8653	2.6296	22.8653
	1.35	0.5106	0.0514	0.5106		3.10	25.3151	2.7314	25.3149

Table 7.4: OC parameter values for γ_0^{OC} and β_0^{OC}, according to each ZIP and ZINB IC scenario.

8 Performance of regression-based control charts

The performance of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart are evaluated based upon their ARL_0, $SDRL_0$, ARL_1 and $SDRL_1$ in this chapter. Each control chart is constructed to achieve an ARL_0 value of 200. The performance is evaluated by simulation of OC data as described in Section 7.6.1, and for each IC scenario. It explained in Section 7.4 that the performance of each chart is also evaluated under two circumstances. At fist, we evaluate the baseline performance of each chart in Section 8.1. Then, the performance while taking into account the effects of Phase I estimation are presented afterwards in Section 8.2.

8.1 Baseline performance of regression-based Shewhart charts

The baseline performance of each (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart is discussed in the following section. The methodology for this performance evaluation is described in Section 7.4.2. In order to keep this chapter structured, it is decided to move the results of IC scenarios 3 and 4 to Appendix B.3. The results for ZIP scenarios 1 and 2 illustrate similar behaviour as for ZINB scenarios 3 and 4 respectively. The obtained values for ARL_1 and $SDRL_1$ are provided in Appendix B.5.

8.1.1 Baseline performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart

The baseline performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart charts with symmetric control limits is shown in Figures 8.1, 8.2, B.13 and B.14, for IC scenarios 1, 2, 3 and 4, respectively. Every figure throughout this chapter contains four out-of-control scenarios. In the upper left corner, we evaluate the performance of each regression-based Shewhart chart, for an OC scenario with increasing $E[\lambda^{OC}]$. An increased expected amount of defects, is considered as worse process performance in the context of monitoring defects in plastic pellet production. In the upper right corner, we evaluate the performance of each chart, for an OC scenario with decreasing $E[p^{OC}]$. A decreased amount of zero-inflation leads to more overall defects as well, such that this is also considered as worse process performance. Besides evaluating the ARL_1 for worse process performance, we also evaluate the ARL_1 in case of improved process performance. The results are shown in the lower two graphs of each figure, whereas the left graph reflects the ARL_1 results for decreasing $E[\lambda^{OC}]$, and the right graph reflects the ARL_1 results for increasing $E[p^{OC}]$.

The first ARL_1 computation is simulated with the IC parameters of the Phase I scenario. Hence, the most left simulation result of each graph reflect the ARL_0 value. When proceeding to the right in each graph, more severe cases of OC data distributions are simulated.

Let us first consider baseline performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart, for an OC scenario due to increased $E[\lambda^{OC}]$. These results are shown in Figures 8.1a, 8.2a, B.13a and B.14a. It is observed from the results that all Shewhart charts with Pearson, deviance and randomised quantile

residuals shows a steep decrease of the ARL_1 as the $E[\lambda^{OC}]$ increases. The (r^D, L)-Shewhart with deviance residuals shows the fastest decrease in ARL_1 in all scenarios. When comparing Figure 8.1a with Figure 8.2a it is observed that the ALR_1 decreases slightly faster in the less zero-inflated scenarios, i.e., ZIP Scenario 2. This conclusion also holds for the ZINB regression-based Shewhart charts of IC scenarios 3 and 4.

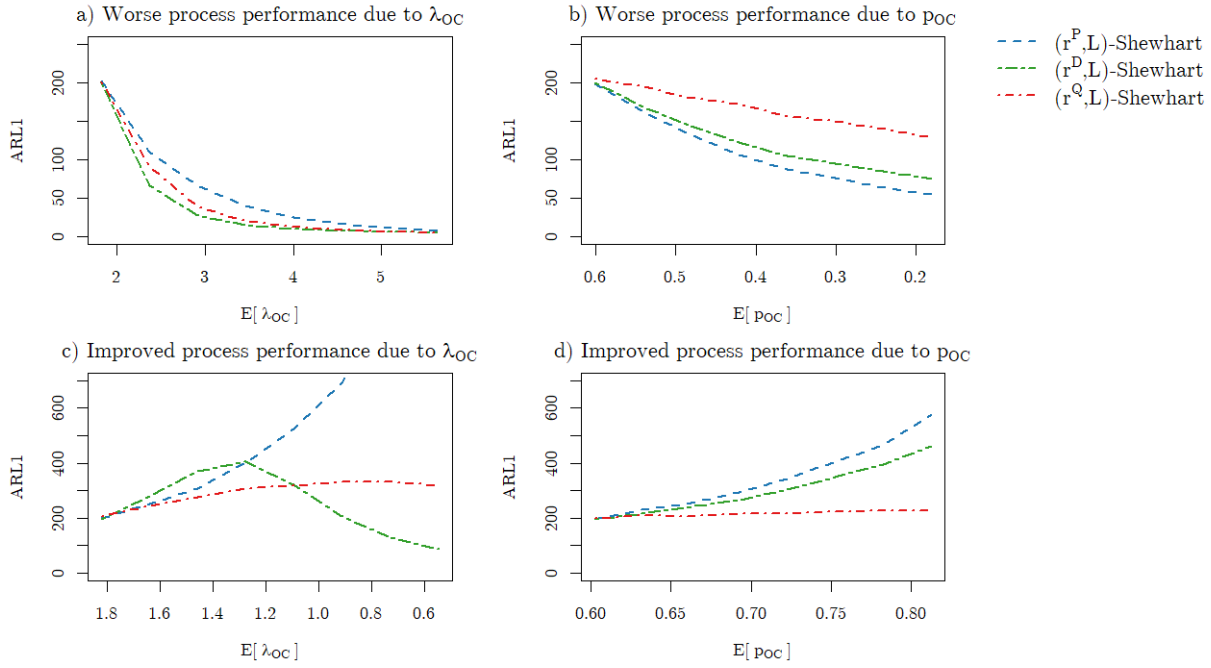

Figure 8.1: Baseline ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

Baseline performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart, for an OC scenario due to decreased $E[p^{OC}]$, is shown in Figures 8.1b, 8.2b, B.13b and B.14b, for IC scenarios 1, 2, 3 and 4, respectively. This OC shift corresponds to worse overall process performance due to less zero-inflation. It is observed that the ARL_1 in each of the charts is slowly decreasing with $E[p^{OC}]$. The decay in ARL_1 is faster for the more zero-inflated data scenarios, i.e., ZIP Scenario 1 and ZINB Scenario 3. In addition, it is observed the charts with Pearson and deviance residuals outperform the chart with randomised quantile residuals in an OC scenario due to decreased $E[p^{OC}]$. The (r^Q, L)-Shewhart chart with randomised quantile residuals does not even detect the OC shift for IC ZIP Scenario 2 and ZINB Scenario 4. This is concluded since the ARL_1 remains stable at 200 with a decrease in $E[p^{OC}]$. From these results it is concluded that Pearson and deviance residuals have best performance in detecting deteriorating process performance.

An improvement of process performance is less well detected in the (r^P, L)-, (r^D, L)-, (r^Q, L)-Shewhart charts. Figures 8.1c, 8.2c, B.13c and B.14c show the baseline ARL_1 performance of each chart under an OC distributional shift due to decreased $E[\lambda^{OC}]$. This OC distributional shift is eventually detected by the (r^D, L)-Shewhart chart with deviance residuals, in case of ZIP distributed data in scenarios 1 and 2. However, the ARL_1 increases first before it decreases, such that small decreases in $E[\lambda^{OC}]$ remain unnoticed. The charts wit Pearson and randomised quantile residuals do not detect the OC shifts. A decrease in $E[\lambda^{OC}]$ remains also unnoticed by all charts in ZINB scenarios 3 and 4.

Finally, Figures 8.1d, 8.2d, B.13d and B.14d show baseline performance of the (r^P, L)-, (r^D, L)-, (r^Q, L)-Shewhart charts, in case of an OC distributional shift due to increased $E[p^{OC}]$. This OC shift corresponds to improved overall process performance due to an increased proportion zero-inflation. These

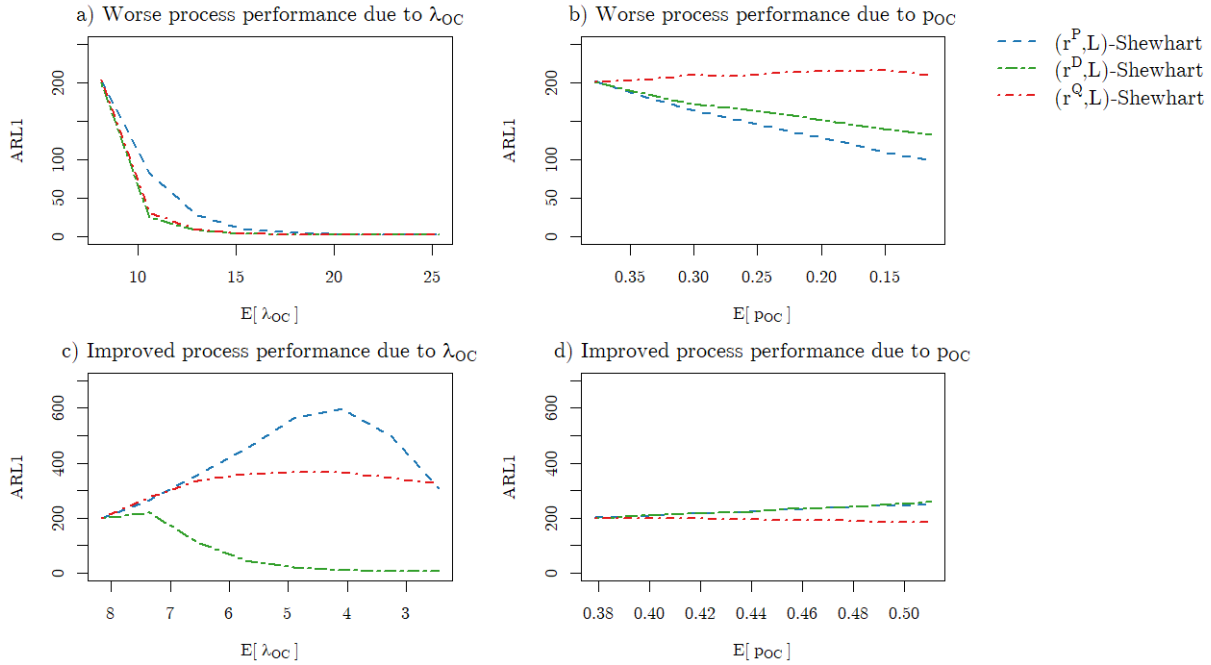

Figure 8.2: Baseline ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart for IC ZIP Scenario 2 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

OC scenarios remain unnoticed by all charts with Pearson, deviance and randomised quantile residuals, since the ARL_1 in only increasing with $E[p^{OC}]$.

All obtained ARL and $SDRL$ values are reported in Appendix B.5. The $SDRL_0$ and $SDRL_1$ results are approximately equal to the corresponding ARL_0 and ARL_1 results, for all baseline performance results for Shewhart charts with symmetric control limits. This is also illustrated in Table 8.1, in which a fraction of all ARL and $SDRL$ results are shown.

This concludes all baseline performance results of regression-based Shewhart charts with symmetric control limits. Baseline performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart are discussed in the next section.

		OC scenario due to increased $E[\lambda^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL_1$	ARL_1	$SDRL_1$	ARL_1	$SDRL_1$
0.10	1.82	202.33	202.66	199.73	198.15	201.62	203.68
0.36	2.37	110.18	108.96	66.51	65.56	91.50	92.01
0.57	2.92	66.61	66.17	27.64	27.32	38.32	37.16
0.74	3.46	39.84	39.62	14.61	14.37	19.52	18.96
0.89	4.01	25.29	25.06	9.37	8.99	11.80	11.46
1.02	4.56	15.92	15.20	7.07	6.48	8.53	7.91
1.13	5.10	10.59	10.02	5.58	5.01	6.47	5.92
1.23	5.65	7.82	7.29	4.74	4.16	5.42	4.81

Table 8.1: Fraction of the baseline ARL_1 results with corresponding $SDRL_1$, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, for IC ZIP Scenario 1.

8.1.2 Baseline performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart

The baseline ARL_1 results of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart with probability control limits are shown in Figures 8.3, 8.4, B.15 and B.16, respectively. All three charts with Pearson, deviance and quantile residuals show a fast decreasing ARL_1 in case of an OC scenario that is due to increased $E[\lambda^{OC}]$. These results are shown in Figures 8.3a, 8.4a, B.15a and B.16a, for IC scenarios 1, 2, 3 and 4, respectively. It is observed that the (r^D, Q)- and (r^Q, Q)-Shewhart chart with deviance and quantile residuals show best performance for all IC scenarios 1, 2, 3 and 4. Nevertheless, the Shewhart charts with symmetric control limits are slightly faster in detecting the OC distributional shift due to increased $E[\lambda^{OC}]$ than the charts with probability control limits, in case of Pearson and deviance residuals. This is due to the fact that both ZIP and ZINB Pearson and deviance residuals have positively skewed distributions.

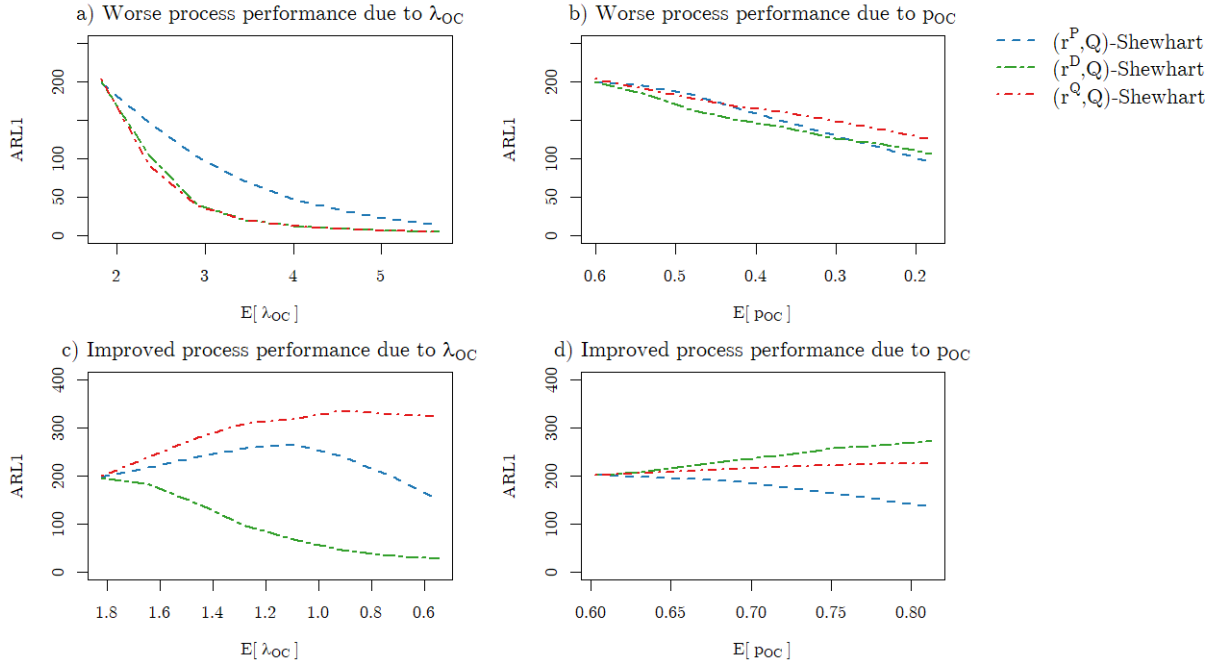

Figure 8.3: Baseline ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

The main difference in performance between the Shewhart charts with probability and symmetric control limits observed in the lower left corner of each plot. Namely, Figures 8.3c, 8.4c, B.15c and B.16c show the performance of (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart in case of an OC distributional shift due to decreased $E[\lambda^{OC}]$. It is observed from the results that the (r^D, Q)-Shewhart chart with deviance residuals shows decreasing ARL_1, for decreasing $E[\lambda^{OC}]$. Hence, it can be concluded from this that the Shewhart chart with probability control limits and deviance residuals is better in detecting process improvement due to a shift in $E[\lambda^{OC}]$ then a similar chart with symmetric control limits. The (r^D, Q)-Shewhart chart with deviance residuals chart also outperforms the (r^P, Q)- and (r^Q, Q)-Shewhart charts with Pearson and randomised quantile residuals, which are not detecting the shift.

The Shewhart charts with probability limits show similar results as the charts wit symmetric control limits, in case of an OC scenarios due to increased or decreased $E[p^{OC}]$. Namely, it is observed from Figures 8.3b, 8.4b, B.15b and B.16b that the ARL_1 of each chart is slowly decreasing with with decreased $E[p^{OC}]$. The (r^D, Q)-Shewhart chart with deviance residuals shows fastest decrease in this case. Figures 8.3d, 8.4d, B.15d and B.16d show the simulated ARL_1 results for OC distributional shifts is due increased $E[p^{OC}]$. These OC shifts remain undetected by the (r^D, Q)- and (r^Q, Q)-Shewhart chart with deviance and randomised quantile residuals. The ARL_1 of the (r^P, Q)-Shewhart chart with Pearson

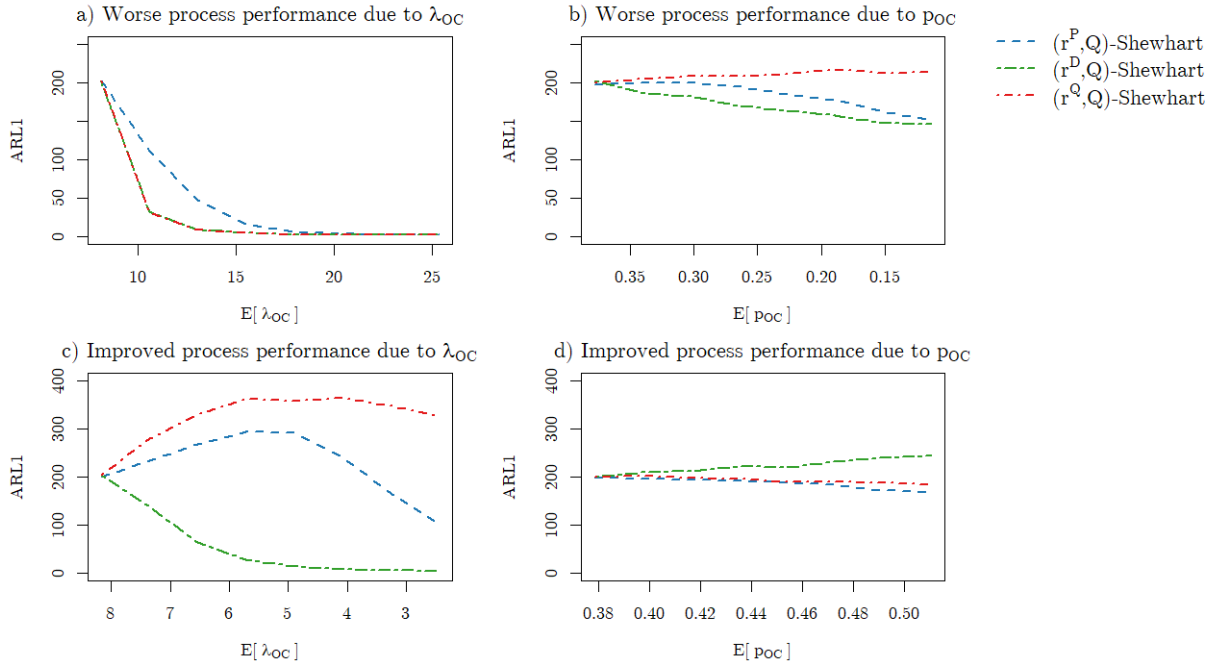

Figure 8.4: Baseline ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart for IC ZIP Scenario 2 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

residuals is only slightly decreasing in case of an IC scenario with high proportion of zero inflation. Hence, in ZIP Scenario 1 and ZINB Scenario 3.

In addition, it is again observed from the results in Appendix B.5, that the $SDRL_0$ and $SDRL_1$ results are approximately equal to the corresponding ARL_0 and ARL_1 results, respectively. It is described in Section 4.2.1 that the run length distribution of a Shewhart chart with symmetric or probability control limits follows a Geometric distribution, in case observations are independent and identically distributed. In that case, $ARL_0 \approx SDRL_0$, which corresponds to the results of the ZIP and ZINB regression-based Shewhart charts. Figure 8.5 shows the run length distribution of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart charts for IC Scenario 1, together with the distribution of the Geometric(0.005) distribution. It is observed that the run length distribution of each charts is very similar to the Geometric(0.005) distribution, even though independence and identical distribution of Pearson, deviance and randomised quantile residuals is not proved.

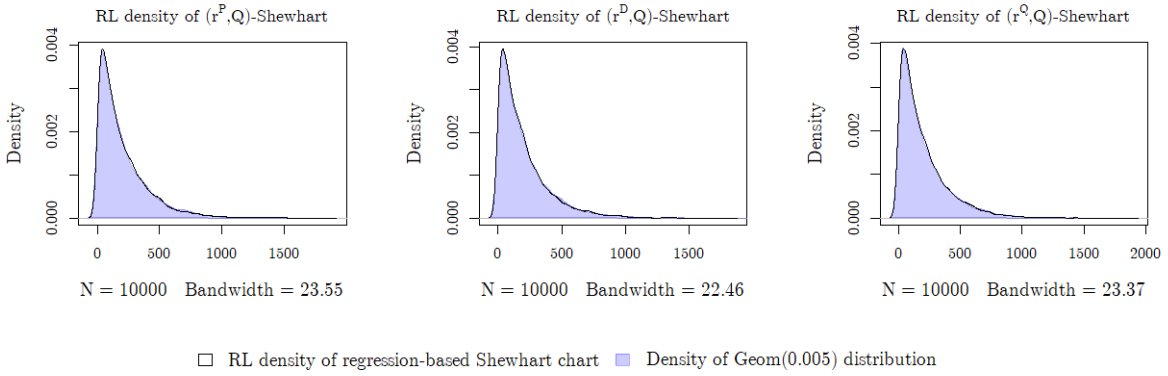

Figure 8.5: IC run length distributions of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart, in ZIP Scenario 1.

This concludes the results on the baseline performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart. In the following section, simulation results are discussed for all control charts while taking into account the effect of Phase I estimation.

8.2 Performance results while considering Phase I effects

In this section, we discuss the simulated ARL_0, $SDRL_0$, ARL_1 and $SDRL_1$ results of the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart, while taking into account the effect of Phase I estimation. These results are obtained according to the methodology described in Section 7.4.2. It is again decided to move the results of IC scenarios 3 and 4 to Appendix B.4, in order to keep this chapter structured. The obtained values of ARL_1 and $SDRL_1$ are provided in Appendix B.5.

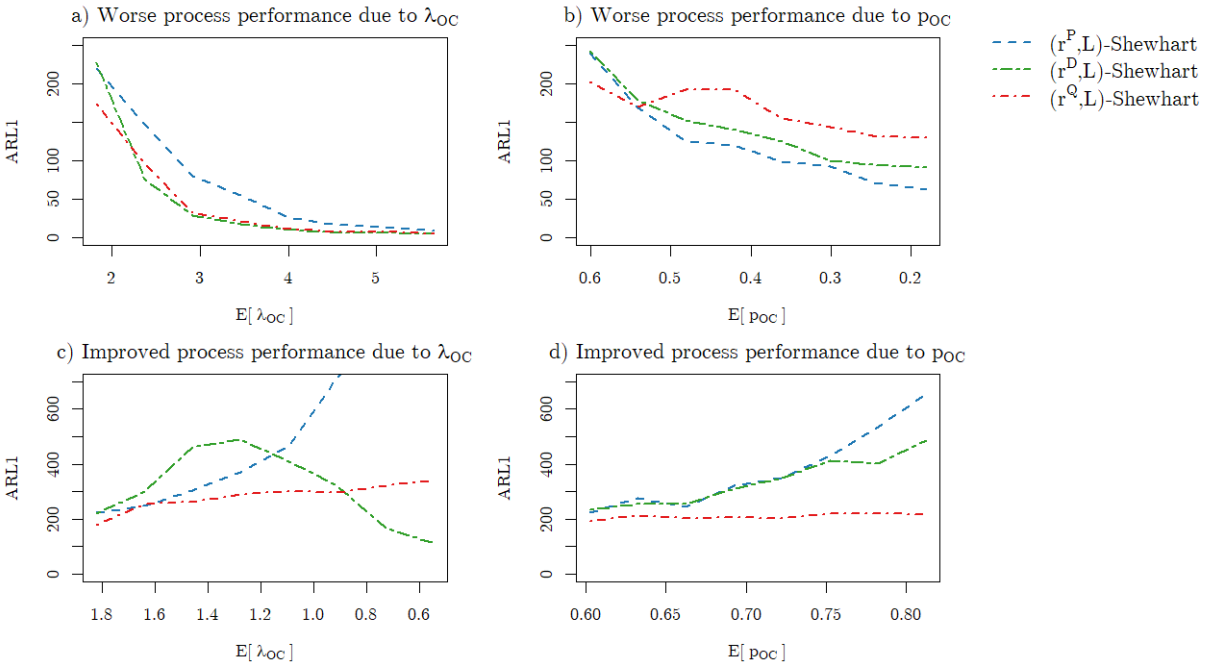

Figure 8.6: ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart, while considering Phase I estimates, for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

Figures 8.6, 8.7, B.20 and B.21 show the simulated ARL_1 results of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart charts, while taking into account the effects of Phase I estimation. These figures are generated for IC ZIP and ZINB scenarios 1, 2, 3 and 4, respectively. First of all, it is noticed that the ARL results in each figure resemble the baseline performance results in Figures 8.1, 8.2, B.13 and B.14, even though ARL behaviour is less smooth. The ARL_0 in each figure is close to 200 in each plot, which shows that, on average, the true ARL_0 is approximately equal to the prespecified ARL_0.

Figures 8.8, 8.9, B.22 and B.23 show obtained ARL_1 results of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart charts with probability control limits, while taking into account the effects of Phase I estimation. It is again observed that the results broadly resemble the baseline performance results, and the true ARL_0 of these charts is on average equal to the prespecified ARL_0 of 200. Baseline performance results of the Shewhart charts with probability control limits are shown in Figures 8.3, 8.4, B.15 and B.16, for IC scenarios 1, 2, 3 and 4, respectively. However, ARL behaviour while taking into account Phase I estimations is again less smooth that what is observed from the baseline performance.

Therefore, a critical not should be made on the performance results that are presented in this section. It is explained in Section 7.5.2 that the $SDRL_0$ of a ZIP and ZINB regression based Shewhart chart

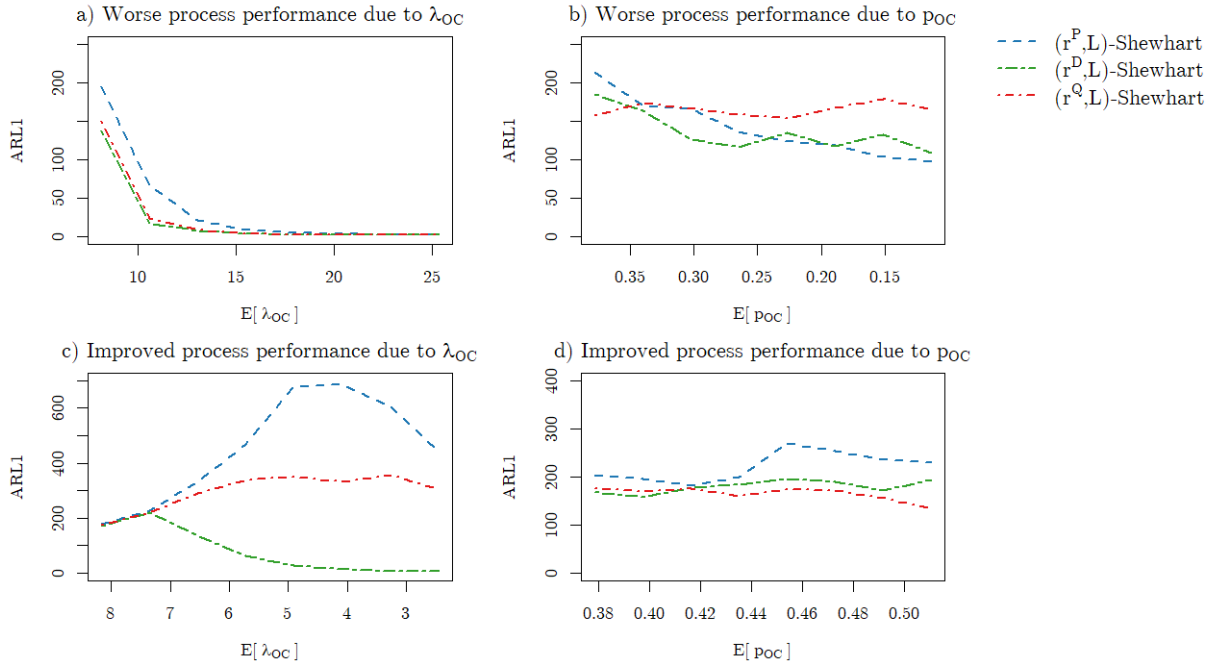

Figure 8.7: ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart, while considering Phase I estimates, for IC ZIP Scenario 2 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

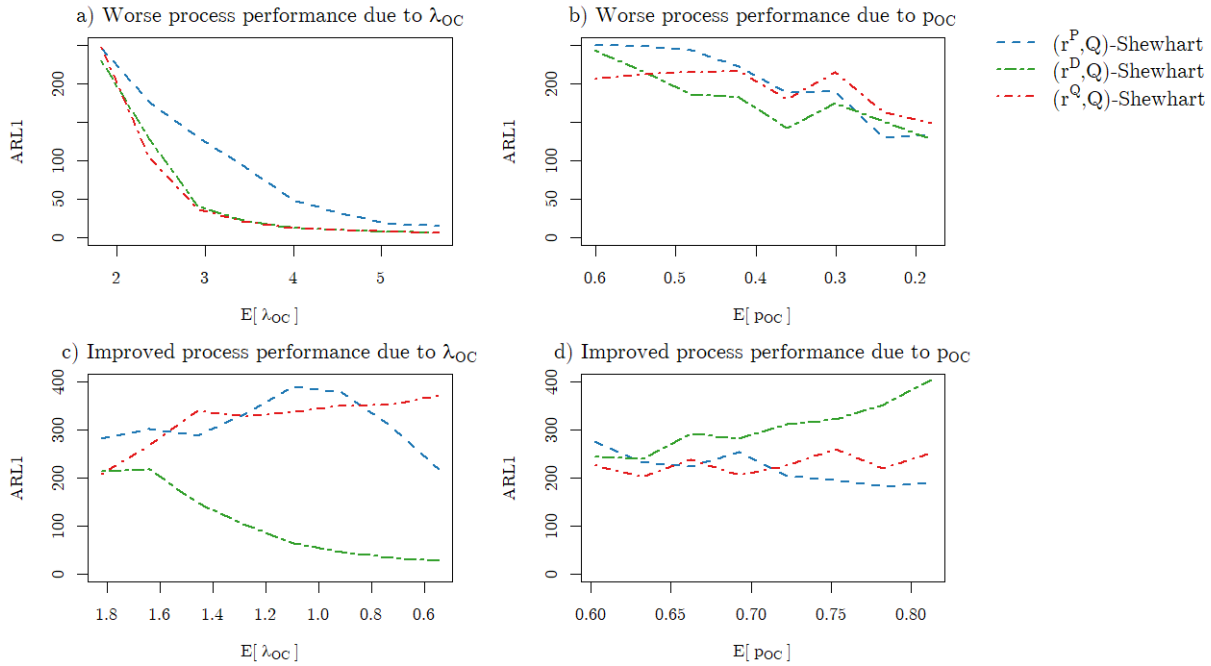

Figure 8.8: ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart, while considering Phase I estimates, for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

stabilises when the number of replications N is large enough. Simulations are executed to show that N=10,000 achieves stable $SDRL_0$ results for Shewhart charts with Pearson, deviance and randomised quantile residuals. Figure 7.8 shows that 100 replications of run length simulations are executed for this performance evaluation. It was decided to take $N = 200$ for each run length simulation while taking into account the effect of Phase I estimation, to ensure an overall simulation size of 20,000 replications.

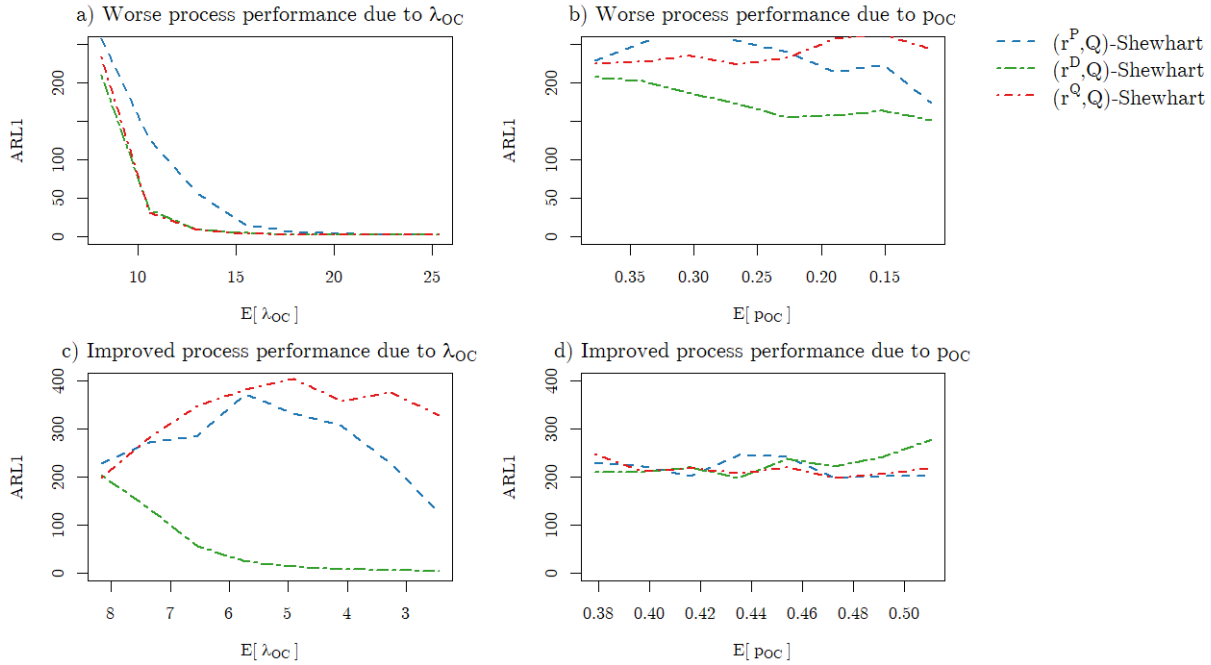

Figure 8.9: ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart, while considering Phase I estimates, for IC ZIP Scenario 2 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

However, this has been an misjudgement. The variation that is inherited in the Phase I data is much larger than anticipated, such that the ARL results are still subject to variation as well. Hence, a larger simulation size N should have been chosen for this performance evaluation strategy, but this has not been possible in this thesis due to time limitations. The ARL results are therefore only an indication of the performance of each regression-based Shewhart chart while taking into account Phase I estimation, but more precise results could be obtained.

In addition to that, it is observed that the SDRL results are even more affected by the small simulation size inside each iteration. Ideally, we would like to obtain the pooled standard deviation over the 100 iterations, where we assume that group averages between each iteration, i.e., ARL results, are not equal. However, in order to obtain the pooled standard deviation, it is necessary to have reliable estimates of the standard deviation in each group. Hence, $N = 10{,}000$ replications should be simulated in each of the 100 iterations. Currently it takes approximately 6 hours to execute the performance evaluation simulations of Pearson, deviance and randomised quantile residuals in one IC scenario, and for one type of control limits. Eight of these simulations have been executed to provide all results in this section. Increasing simulation size N from 200 to 10,000 is expected to increase the computation time heavily, whereas it might take over 300 hours per simulation. Nevertheless, it is recommended for future research at Dow or in general to use a simulation size of 10,000 replications per Phase I iteration anyway. Only then, the true pooled standard deviation is can be analysed, which would provide valuable information regarding the run length variability due to Phase I estimation.

Table 8.2 shows the pooled standard deviations for a fraction of the current simulation results, which are denoted with SD^p_{RL}. Overall standard deviations that were calculated over all 20,000 observations are provided in Table 8.2 as well, which are denoted with SD_{RL}. The pooled standard deviation is approximately ten times the size of the corresponding ARL, while the overall standard deviation is approximately equal to the ARL. Hence, no conclusions can be made from these results, such that SDRL is not futher considered as a performance measure in this section.

		OC scenario due to increased $E[\lambda^{OC}]$								
		(r^P, Q)-Shewhart			(r^D, Q)-Shewhart			(r^Q, Q)-Shewhart		
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	SD_{RL}	SD_{RL}^p	ARL_1	SD_{RL}	SD_{RL}^p	ARL_1	SD_{RL}	SD_{RL}^p
0.10	1.82	281.56	278.64	1986.40	214.37	207.26	1954.34	207.58	214.03	2004.01
-0.01	1.64	301.06	263.38	2250.28	218.41	233.30	1784.37	267.05	265.05	2396.26
-0.12	1.46	288.76	253.25	2441.75	148.56	162.80	1414.13	340.44	327.55	2742.55
-0.26	1.28	332.81	356.67	2601.42	101.92	102.02	981.10	327.77	335.40	3068.97
-0.41	1.09	389.38	397.80	2695.21	64.69	71.78	673.07	338.09	327.47	3166.75
-0.59	0.91	378.08	388.41	2578.85	45.34	50.64	467.34	350.15	368.48	3282.43
-0.82	0.73	307.22	358.86	2155.20	34.34	33.62	343.23	352.23	303.41	3258.06
-1.10	0.55	217.13	205.67	1600.31	27.72	27.79	279.71	371.50	350.74	3186.45

Table 8.2: Fraction of the ARL_1 results with corresponding SD_{RL} and SD_{RL}^p while taking into account Phase I estimation effects, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, in IC ZIP Scenario 1.

8.3 Summary

Simulation results from the (r^P, L)-, (r^D, L)-, (r^Q, L)-, (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart are discussed in this section. It is observed that the (r^D, Q)-Shewhart charts with deviance residuals and probability control limits outperforms all other charts, in case of an OC distributional shift in $E[\lambda^{OC}]$. This is concluded from both baseline performance, as from performance analysis while taking into account Phase I effects. It also holds for each IC ZIP and ZINB scenarios 1, 2, 3 and 4.

OC scenarios due to distributional shift in $E[p^{OC}]$ are less well detected by all Shewhart charts. The Shewhart charts with Pearson and deviance residuals show a slow decreasing ARL_1 in case of decreased $E[p^{OC}]$. However, none of the charts show convincing decreased ARL_1 results, in case of OC distributional shift due to increased $E[p^{OC}]$. This is observed in the baseline performance, as well as in performance results which take into account Phase I effects.

It is additionally observed that each control chart performs better under less zero-inflated IC circumstances. Hence, all ARL_1 results decrease faster in ZIP Scenario 2 than in ZIP Scenario 1. Similarly, ARL_1 results decrease faster in ZIP Scenario 4 than in ZIP Scenario 3. However, these differences are very modest. Finally, it is noted that the simulation size should be increased, for performance evaluation while estimating Phase I effects. Finally, it is observed that the run length distribution of Shewhart charts with Pearson, deviance and randomised quantile residual is very similar to a Geometric(0.005) distribution, even though independence and identical distribution of Pearson, deviance and randomised quantile residuals is not proved.

9 | GLM-based TBE charts

It is concluded in the previous chapter that the ZIP and ZINB regression-based Shewhart charts work satisfactory for detecting contextual anomalies in monitoring data that originates from high-purity processes. However, it is observed that an OC scenario due to increased or decreased $E[p^{OC}]$ is hardly detected and the overall performance of each charts reduces slightly for larger proportions of zero-inflation in the IC data. The latter may create a problem for some very high-purity processes at Dow with heavily zero-inflated monitoring data. Therefore we explore an alternative GLM-based monitoring method in this chapter.

It is described in Rizzo et al. (2020) that time-between-events (TBE) charts are appropriate tools for monitoring such very high-purity processes, because they overcome the methodological challenges that standard control charts present due to extreme proportions of zero-inflation. TBE control charts originate from Calvin (1983), where it is proposed to monitor the conforming run length (CRL) of low-defect rate processes, instead of monitoring each individual observation. Based on this idea, Goh (1987) introduced cumulative count control chart (CCC) for discrete time observations. Xie et al. (2002b) introduced the t_r-chart for continuous time observation from a homogeneous Poisson process. Here, the time until the rth nonconforming event is monitored over time. In this chapter, we introduce a GLM-based TBE chart for monitoring data from high-purity processes. A TBE data description is provided in the following section, after which the Gamma GLM is defined in Section 9.2. The monitoring procedure of Gamma GLM-residuals is described in Section 9.3, after which the procedure for OC performance evaluation is defined in Section 9.4. Performance results are finally presented Section 9.5

9.1 TBE data description

In the context of monitoring plastic pellets for defects, we can monitor the time until the rth non-zero occurrence, i.e. $Y_i > 0$ for $i = 1, 2, \ldots$. It is described in Chapter 2 that plastic pellet production is a continuous process, such that a portion of pellets is inspected continuously on minute basis. Hence, observations (Y_i, X_i) with $i = 1, 2, \ldots$ arrive every minute, where Y_i represents the detected defect count and X_i represents the inspected weight. We can consider the indicator function $\mathbb{1}_{Y_i>0}$ for $i = 1, 2, \ldots$ as a non-homogeneous Bernoulli process. If Y_i follows a ZIP(p_i,λ_i) distribution, then

$$P(\mathbb{1}_{Y_i>0} = v) = \begin{cases} p_i + (1-p_i)e^{-\lambda_i} & \text{if } v = 0 \\ 1 - p_i - (1-p_i)e^{-\lambda_i} & \text{if } v = 1 \end{cases}$$

as defined in (6.7). If Y_i follows a ZINB(p_i,λ_i,τ) distribution, then

$$P(\mathbb{1}_{Y_i>0} = v) = \begin{cases} p_i + (1-p_i)(1+\lambda_i/\tau)^{-\tau} & \text{if } v = 0 \\ 1 - p_i - (1-p_i)(1+\lambda_i/\tau)^{-\tau} & \text{if } v = 1 \end{cases}$$

as defined in (6.13). In both cases, parameters p_i and λ_i depend on covariate X_i such that the Bernoulli probability is not constant for $i = 1, 2, \ldots$. Instead of monitoring observations (Y_i, X_i) over time, we can monitor the time until the rth nonconforming event, i.e., $Y_i > 0$, while taking into account the accumulated inspected weight. Let us denote the time until the rth non-zero occurrence with T_j and the corresponding accumulated inspected weight with W_j. For a given data set $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$, the TBE data can be obtained as $(t, w) = \{(t_1, w_1), \ldots, (t_h, w_h)\}$ where $h = \sum_{i=1}^{m} \lfloor \mathbb{1}_{\{y_i>0\}}/r \rfloor$. Figure 9.1 shows a graphical representation of this transformation.

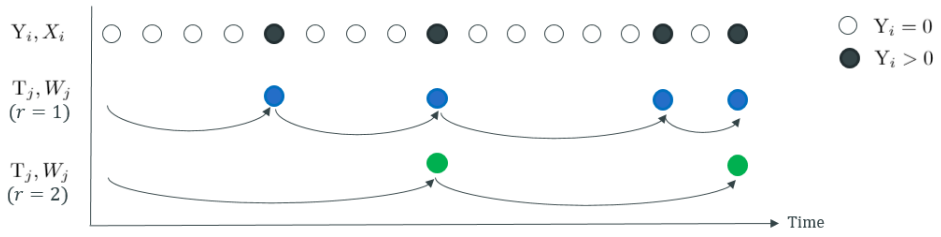

Figure 9.1: Graphical representation of the Bernoulli process $\mathbb{1}_{Y_i>0}$ with $i = 1, 2, \ldots$ and corresponding time between events T_j with accumulated inspected weights W_j with $j = 1, 2, \ldots$.

A GLM-based TBE chart is constructed in this chapter, where we correct for the effect that W_j has on T_j. Observations Y_i follow a ZIP(p_i, λ_i) or ZINB(p_i, λ_i, τ) distribution, of which the parameters p_i and λ_i depend on X_i. Hence, observations Y_i are not independent, nor identically distributed. The distribution of T_j remains therefore unknown. Nevertheless, we can analyse its distribution by simulation. Let us simulate 1500 ZIP and ZINB distributed observations as defined in (7.8), and for each IC scenario. Histograms of the corresponding time-between-events variable T_j are shown in Figure 9.2. It is observed that the distribution of T_j is skewed for every IC scenario, and that the average value of T_j is higher for the IC scenarios that contain a large proportion of zero-inflation, i.e. ZIP Scenario 1 and ZINB Scenario 3.

In addition, we can fit several EDM distributions to the corresponding TBE data T_j, and evaluate the log-likelihood as a goodness of fit measure. Log-likelihood results are shown in Table 9.1 where the Normal, Poisson, negative binomial, exponential and Gamma distribution are fitted.

Distribution	ZIP Scenario 1	ZIP scenario 2	ZINB scenario 3	ZINB scenario 4
Normal	-1069.75	-1318.04	-1133.06	-1287.88
Poisson	-1038.96	-1288.34	-1087.26	-1270.66
Negative binomial	-970.48	-1288.93	-1012.41	-1271.47
Exponential	-979.34	-1363.29	-1014.27	-1351.95
Gamma	-933.48	-1055.99	-966.54	-1058.79

Table 9.1: Log-likelihood of distribution fit for T_j, for 1500 simulated observations Y_i according to all IC ZIP and ZINB scenarios.

It is observed that the Gamma distribution achieves highest log-likelihood in all IC scenarios, and goodness of fit increases with higher proportions of zero-inflation in the simulated observations Y_i. We aim to construct the GLM-based TBE chart for high-purity processes with extreme proportions of zero-inflation, such that the Gamma distribution seems to be an appropriate choice. Let us therefore assume for now that T_j also follows a continuous Gamma distribution. We continue under this assumption and explore the performance of the Gamma GLM-based TBE chart in the following sections.

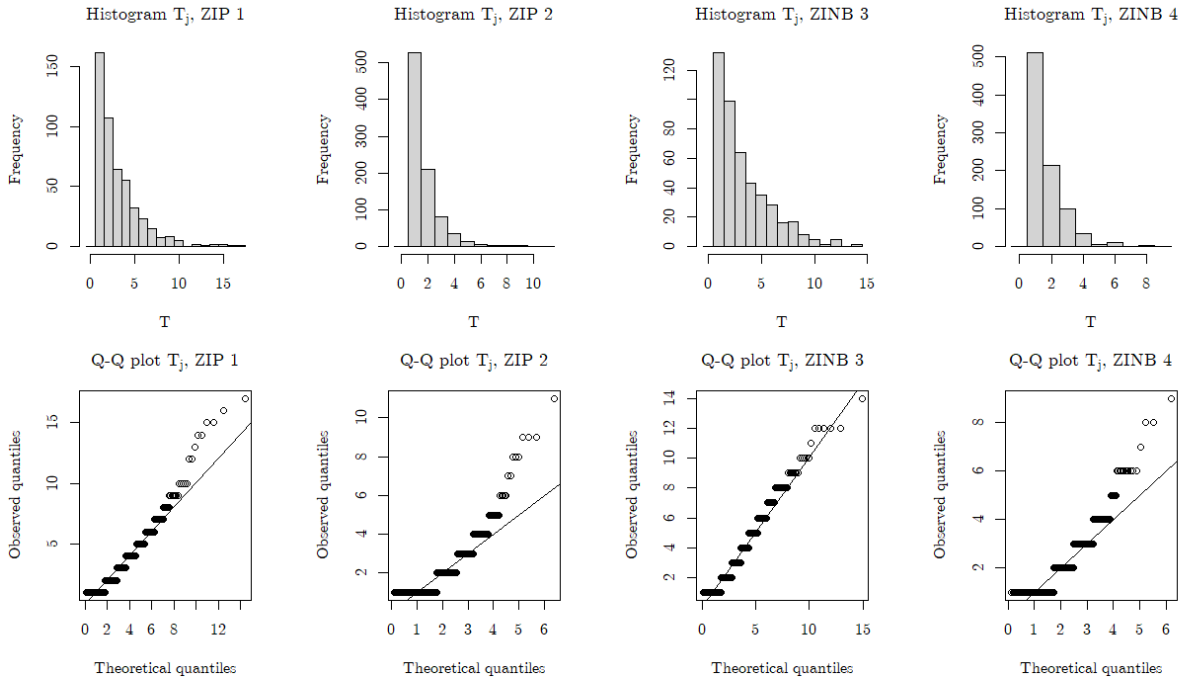

Figure 9.2: Histograms and Gamma Q-Q plots of T_j, for 1500 simulated observations Y_i according to all IC ZIP and ZINB scenarios.

9.2 The Gamma GLM

Let us assume that the time between events T_j follows a Gamma distribution with shape parameter k and scale parameter s_j. The probability distribution function of T_j at point $t_j > 0$ is defined as

$$f_{T_j}(t_j, k, s_j) = \frac{1}{s_j^k \, \Gamma(k)} t_j^{k-1} \exp\{-\frac{t_j}{s_j}\}. \tag{9.1}$$

with $k, s_j > 0$ for all $j = 1, 2, \ldots$. The expected value of T_j is defined as $E[T_j] = ks_j$ of which is provided by Lemma A.1.5. As shown in Lemma A.1.6, the variance of T_j is defined as $\text{Var}(T_j) = ks_j^2$. The Gamma distribution belongs to the exponential family such that a Gamma GLM is defined by Nelder and Wedderburn (1972). Let us persist with the notation of Chapter 6. According to Dunn and Smyth (2018) (Section 11.2), the Gamma distribution can be written in EDM format when we define the EDM components as follows.

- Canonical parameter $\theta_j = -\frac{1}{\mu_j} = -\frac{1}{ks_j}$.
- Cumulant function $\kappa(\theta_j) = -\log\left(\frac{1}{ks_j}\right)$.
- Dispersion parameter $\varphi = \frac{1}{k} > 0$.
- Normalising function $a(t_j, \varphi) = k^k \Gamma(k)^{-1} t_j^{k-1}$.

Hence, substituting these components into (6.1) provides us with the Gamma probability density function, that is shown below.

$$\mathcal{P}(t_j; \theta, \varphi) = a(t_j, \varphi) \exp\left\{\frac{t_j\theta_j - \kappa(\theta_j)}{\varphi}\right\} = k^k \Gamma(k)^{-1} t_j^{k-1} \exp\left\{k\left(-\frac{t_j}{ks_j} + \log\left(\frac{1}{ks_j}\right)\right)\right\}$$

$$= k^k \Gamma(k)^{-1} t_j^{k-1} t_j^{k-1} \exp\{-\frac{t_j}{s_j}\} \left(\frac{1}{ks_j}\right)^k = \frac{1}{s_j^k \, \Gamma(k)} t_j^{k-1} \exp\{-\frac{t_j}{s_j}\}$$

The variance function of the Gamma EDM is defined as $V(\mu_j) = \mu_j^2 = k^2 s_j^2$, where it holds that $\varphi V(\mu_j) = \text{Var}(T_j)$, as described in Section 6.1.1. With this EDM formulation of the Gamma distribution, we can define a Gamma GLM that models the relationship between the time-between-events variable T_j and the accumulated weights W_j. Substituting the EDM Gamma distribution into (6.2) provides us with the following model

$$\begin{cases} f_{T_j}(t_j, k, s_j) = \dfrac{1}{s_j^k \, \Gamma(k)} t_j^{k-1} \exp\{-\frac{t_j}{s_j}\} \\ g(\mu_j) = g(k s_j) = \nu_0 + \nu_1 W_j \end{cases}$$

where ν_0 and ν_1 represent the regression coefficients. The canonical link function of the Gamma GLM is defined as $g(\mu_j) = \theta = -1/\mu_j$. However, Dunn and Smyth (2018) (Section 11.1) state that the logarithmic link function is often used since it avoids the need for constraints on the linear predictor to achieve $\mu_j > 0$. Hence, we proceed in a similar way and define the Gamma GLM as

$$\begin{cases} f_{T_j}(t_j, k, s_j) = \dfrac{1}{s_j^k \, \Gamma(k)} t_j^{k-1} \exp\{-\frac{t_j}{s_j}\} \\ \mu_j = k s_j = \exp\{\nu_0 + \nu_1 W_j\}. \end{cases} \tag{9.2}$$

Now that the Gamma GLM is defined, we can proceed to the following section, in which monitoring of Gamma GLM residuals is described.

9.3 Monitoring Gamma GLM residuals

Let us assume that we have a Phase I data set $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$, from which the Phase I TBE data $(t, w) = \{(t_1, w_1), \ldots, (t_h, w_h)\}$ is obtained with $h = \sum_{i=1}^{m} \lfloor \mathbb{1}_{\{y_i > 0\}} / r \rfloor$. Then, regression coefficients ν_0, ν_1 and k are estimated by maximisation of the log-likelihood function

$$\ell(\nu_0, \nu_1, k; t) = \sum_{j=1}^{h} \log P(t_j; s_j, k)$$

where $s_j = \frac{1}{k} \exp\{\nu_0 + \nu_1 w_j\}$. The maximisation procedure is executed by means of the iterative weighted least squares (IWLS) algorithm. These ML estimates are denoted with $\hat{\nu}_0$, $\hat{\nu}_1$ and $\hat{k}$ respectively, and the model is referred as the fitted Gamma model. Now, let us denote a new observations with (t_{h+j}, w_{h+j}) for $j = 1, 2, \ldots$. We can use the fitted Gamma model to predict the value of t_{h+j}, based on the value of w_{h+j}. The predicted value is denoted with $\hat{\mu}_{h+j}$, which is defined as

$$\hat{\mu}_{h+j} = \hat{k} \hat{s}_{h+j} = \exp\{\hat{\nu}_0 + \hat{\nu}_1 w_{h+j}\}.$$

For each new observation (t_{h+j}, w_{h+j}) with $j = 1, 2, \ldots$ and corresponding prediction $\hat{\mu}_{h+j}$, we can obtain regression residuals and monitor them over time. Namely, Gamma GLM Pearson, deviance and quantile residuals residuals are obtained as described in Section 6.1.4. Pearson residuals for a Gamma GLM are defined by substituting $V(\hat{\mu}_{h+j}) = \hat{k}^2 \hat{s}_{h+j}^2$ into (6.3). Hence, let us denote the Gamma GLM Pearson residuals with r_{h+j}^P where

$$r_{h+j}^P = \frac{t_{h+j} - \hat{\mu}_{h+j}}{\hat{k} \hat{s}_{h+j}} \tag{9.3}$$

for $j = 1, 2, \ldots$. In addition, Gamma GLM deviance residuals are obtained by substituting $\theta = -1/\hat{\mu}_{h+j}$, $\kappa(\theta) = -\log(1/\hat{\mu}_{h+j})$ and $\hat{\varphi} = 1/\hat{k}$ in definition (6.4). When denoting the deviance residuals with r_{h+j}^D, we obtain

$$r_{h+j}^D = \text{sign}(t_{h+j} - \hat{\mu}_{h+j}) \sqrt{2 \left\{ -\log\left(\frac{t_{h+j}}{\hat{\mu}_{h+j}}\right) + \frac{t_{h+j} - \hat{\mu}_{h+j}}{\hat{\mu}_{h+j}} \right\}} \tag{9.4}$$

for $j = 1, 2, \ldots$. Finally, we can define the Gamma GLM quantile residuals by substituting the Gamma cumulative distribution function into (6.6). Let us denote the Gamma GLM quantile residuals with r^Q_{h+j}. Then

$$r^Q_{h+j} = \Phi^{-1}\{F(t_{h+j}; \hat{\mu}_{h+j}/\hat{k}, \hat{k})\} \tag{9.5}$$

where

$$F(t; s, k) = \frac{1}{s^k\,\Gamma(k)} \int_0^t x^{k-1} \exp\{-x/s\} dx$$

for $j = 1, 2, \ldots$. Each type of residuals can be monitored over time in a Gamma GLM-based control chart. When monitoring predictive Pearson, deviance or quantile residuals, it is assumed that observations T_{h+j} with $j = 1, 2, \ldots$ follow a Gamma distribution with parameters $\hat{s}_{h+j}$ and $\hat{k}$ as long as the process is in control. The parameters $\hat{s}_{h+j}$ are predicted based on the observed value of W_{h+j} and the established regression coefficients $\hat{\nu}_0, \hat{\nu}_1$, and $\hat{k}$ from the Phase I GLM that is defined in (9.2). Let us denote this assumption with $T_{h+j} \sim \text{Gamma}(\hat{s}_{h+j}, \hat{k} \,|\, W_{h+j})$. Then the hypotheses of the Gamma GLM-based control chart can be defined as

$$H_0: \quad T_{h+j} \sim \text{Gamma}(\hat{s}_{h+j}, \hat{k} \,|\, W_{h+j}) \quad \text{for} \quad j = 1, 2, \ldots$$

$$H_1: \begin{cases} T_{h+j} \sim \text{Gamma}(\hat{s}_{h+j}, \hat{k} \,|\, W_{h+j}) & \text{for} \quad j = 1, \ldots, \mathcal{T} \\ T_{h+j} \sim \text{Gamma}(s^{OC}_{h+j}, k^{OC} \,|\, W_{h+j}) & \text{for} \quad j = \mathcal{T}+1, \mathcal{T}+2, \ldots. \end{cases}$$

where $s^{OC}_{h+j} \neq \hat{s}_{h+j}$ and or $k^{OC} \neq \hat{k}$. If the process becomes OC after time $\mathcal{T}$, then it is either rejected that the shape parameter equals IC estimation $\hat{k}$, and or it is rejected that the scale parameter equals IC estimation $\hat{s}_{h+j}$ for $j = \mathcal{T}+1, \mathcal{T}+2, \ldots$. In this case, s^{OC}_{h+j} is defined as

$$s^{OC}_{h+j} = \exp\{\nu_0^{OC} + \nu_1^{OC} w_{h+j}\} \tag{9.6}$$

where at least one of the following holds: $\hat{\nu}_0 \neq \nu_0^{OC}$ or $\hat{\nu}_1 \neq \nu_1^{OC}$. Hence, the process becomes OC when observations $\{T_{h+\mathcal{T}+1}, T_{h+\mathcal{T}+2}, \ldots\}$ follow a Gamma distribution with parameters that deviate from what is expected by the established GLM, based on the values of $\{W_{h+\mathcal{T}+1}, W_{h+\mathcal{T}+2}, \ldots\}$.

Each residual type is monitored over time in a GLM-based Shewhart chart. It is concluded from the results in Chapter 8 that Shewhart charts with probability control limits have better overall performance than Shewhart charts with symmetric control limits. Therefore, we construct GLM-based TBE charts with probability control limits, and for Pearson, deviance and quantile residuals. Let us denote each chart with (r^P, Q)-TBE, (r^D, Q)-TBE, and (r^Q, Q)-TBE. The construction of each chart and the strategy for performance evaluation is discussed in the following section.

9.4 OC performance evaluation of the GLM-based TBE chart

The goal of this chapter is to compare the performance of the GLM-based TBE charts with the performance of the ZIP and ZINB regression base Shewhart charts. However, observations in the TBE chart are on a different time scale as the observations in the ZIP and ZINB regression-based Shewhart charts. This is illustrated in Figure 9.1. Suppose that observations (Y_{m+i}, X_{m+i}) with $i = 1, 2, \ldots$ arrive every minute. Then variable T_{h+1} represents the cumulative time until the rth nonconforming event, i.e., $\min_{i=1,2,\ldots}(Y_{m+i} > 0)$. The same reasoning holds for T_{h+j} with $j = 1, 2, \ldots$ such that the arrival rate of observations (T_{h+j}, W_{h+j}) is not constant.

It is stated in Rizzo et al. (2020) that, when the time interval between cumulative observation varies, the average length of inspection (ALI) is a more appropriate performance measure than the ARL. The ALI is also referred to as the average number of observations to signal (ANOS). For a particular run of

Gamma GLM residuals $\{r_{h+1}, \ldots, r_{h+RL}\}$, the length of inspection (LI) is defined as

$$LI = \sum_{j=1}^{RL} T_{h+j} \tag{9.7}$$

where RL is the run length. Hence, the ALI of a Gamma GLM-based TBE chart is on the same time scale as the ARL of a ZIP and ZINB regression-based Shewhart charts, in case observations arrive on minute basis. This allows for performance comparison. The IC ALI is referred to as ALI_0 while the OC ALI is denoted with ALI_1. The control limits of each (r^P, Q)-TBE, (r^D, Q)-TBE, and (r^Q, Q)-TBE chart are constructed to achieve an ALI_0 of 200. The performance of each chart is evaluated based on the ALI_1.

9.4.1 Performance evaluation strategy of GLM-based TBE charts

We will test the performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart, for each ZIP and ZINB IC data scenario defined in Table 7.1. It is assumed that all ZIP and ZINB regression parameters β_0, β_1, γ_0, γ_1 and τ are known, to exclude any variability in performance that originates from estimating ZIP and ZINB regression coefficients. The consecutive steps of solving control limits and evaluating OC performance of the Gamma GLM-based TBE charts are provided in the following steps, of which a graphical representation is shown in Figure 9.3.

1. **Phase I data generation:** a) A Phase I data set of size m is generated with ZIP or ZINB distributed observations $(y, x) = \{(y_1, x_1), \ldots, (y_m, x_m)\}$. Known parameters β_0, β_1, γ_0, γ_1 and τ are applied in the simulation that is defined in (7.8). b) The Phase I data set (y, x) is transformed to obtain the time-between-events and accumulated weights between events. Hence, a data set $(t, w) = \{(t_1, w_1), \ldots, (t_h, w_h)\}$ is obtained with $h = \sum_{i=1}^{m} \lfloor \mathbb{1}_{\{y_i > 0\}} / r \rfloor$.

2. **Fit the Phase I Gamma GLM:** The Gamma GLM as defined in (9.2) is fitted to the Phase I TBE data (t, w), to obtain estimates $\hat{\nu}_0$, $\hat{\nu}_1$ and $\hat{k}$.

3. **Constructing TBE control limits:** a) An IC Phase II data set is generated with N runs of n ZIP or ZINB distributed observations. Let us denote these observations with $(y_{\ell,m+i}, x_{\ell,m+i})$ where $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$. Known parameters β_0, β_1, γ_0, γ_1 and τ are applied in the simulation, which is defined in (7.8). b) The IC Phase II data set is transformed to obtain the TBE data, whereas the length of each TBE run equals the amount of non-zero observations in the corresponding ZIP or ZINB data run. Let us define $n_\ell = \sum_{i=m+1}^{m+n} \mathbb{1}_{\{y_{\ell,m+i} > 0\}}$ for each run of observations $\ell = 1, \ldots, N$. Then, the IC TBE Phase II data set is defined as $(t_{\ell,h+j}, w_{\ell,h+j})$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. c) The Phase I Gamma GLM is applied to obtain the regression residuals from each observation $(t_{\ell,h+j}, w_{\ell,h+j})$. These regression residuals are denoted with $r_{\ell,h+j}$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. The control limits of the GLM-based TBE chart Q_1 and Q_2 are constructed to achieve an ALI_0 of 200, as described in Section 9.4.2.

4. ALI_1 **performance evaluation:** a) Once the UCL and LCL of the Gamma GLM-based TBE chart are defined, an OC Phase II data set is generated with again N runs of n ZIP or ZINB distributed observations. The first simulated data point is already OC, such that $m = \mathcal{T}$ where $\mathcal{T}$ is the changepoint and m denotes the size of the Phase I data. Observations are denoted with $(y_{\ell,\mathcal{T}+i}, x_{\ell,\mathcal{T}+i})$ with $i = 1, \ldots, n$ and $\ell = 1, \ldots, N$. The OC parameters $\beta_0^O C$, $\beta_1^O C$, $\gamma_0^O C$, $\gamma_1^O C$ and τ are applied in the simulation, which is described in Section 7.6. b) The OC Phase II data set is transformed to obtain the OC TBE data. Now we have, $n_\ell = \sum_{i=\mathcal{T}+1}^{\mathcal{T}+n} \mathbb{1}_{\{y_{\ell,m+i} > 0\}}$ and $(t_{\ell,h+j}, w_{\ell,h+j})$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. c) The Phase I Gamma GLM applied to obtain the regression residuals from each observation $(t_{\ell,h+j}, w_{\ell,h+j})$. These regression residuals

are denoted with $r_{\ell,h+j}$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. The run length RL_ℓ of each residual run l is obtained according to the values of Q_1 and Q_2, that are defined in Step 3. The LI_ℓ of each run is obtained afterwards, according to definition (9.7). Finally, the ALI_1 is obtained as the mean of all LI_ℓ values with $\ell = 1, \ldots, N$.

5. Step 1-4 are repeated 100 times.

Steps 1-4 are repeated in this strategy to eliminate the variation that is inherited in the Phase I data. The ALI results from step 4 are accumulated and averaged over all 100 iterations, to obtain the final results. Size parameters are again chosen at $m = 1500$, $n = 3000$ and $N = 200$ as described in Section 7.5.2. It is analysed again whether simulation size n is large enough to ensure $\leq 0.01\%$ of the runs returns no OC signal. The simulation results of this analysis are provided in B.20, from which it is concluded that $n = 3000$ is sufficiently large. Similar limitations as described in Section 8.2 are experienced due to the choice of $N = 200$. Therefore, we do not consider the $SDLI$ as a performance measure in this chapter. Aggregation level r is fixed at $r = 1$ for all simulations and performance evaluations.

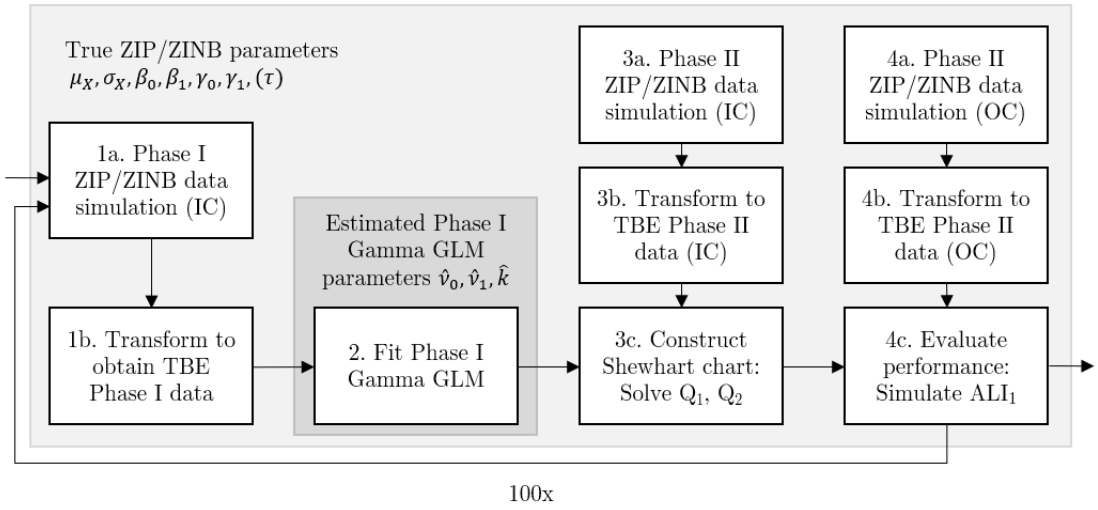

Figure 9.3: Graphical representation of performance evaluation strategy of the Gamma GLM-based TBE chart.

Notice that is not possible to establish a baseline performance for the GLM-based TBE charts, since the distributions of T_j and W_j are unknown. Therefore, the true relation between T_j and W_j is also unknown. A Gamma GLM is fitted nevertheless, but $\hat{\nu}_0$, $\hat{\nu}_1$ and $\hat{k}$ cannot be assumed to capture the true regression model. Hence, baseline performance is not obtained, but an estimation of the Gamma GLM-based TBE chart performance is obtained while taking Phase I estimation effects of the GLM into account and under the assumption that all ZIP or ZINB regression parameters are known. The performance results in terms of ALI_1 are provided in the following Section 9.5. However, it is described first how the probability control limits of the GLM-based TBE charts are obtained to achieve $ALI_0 = 200$.

9.4.2 Obtaining probability limits for the GLM-based TBE chart

Notice that we cannot solve the probability limits as described in Section 7.5.1 for the ZIP and ZINB regression-based Shewhart charts, since we are not interested in the ARL of the TBE charts. Instead, we must obtain a quantile level α of the probability control limits, that ensures $ALI_0 = 200$. Let us continue with the notation in Step 3 of the previous section. Hence, let us denote the IC TBE Phase II data set with $(t_{\ell,h+j}, w_{\ell,h+j})$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. Here, $n_\ell = \sum_{i=m+1}^{m+n} \mathbb{1}_{\{y_{\ell,m+i}>0\}}$ for

all runs $\ell = 1, \ldots, N$. The following steps are executed to solve the probability limits of a GLM-based TBE chart:

1. The Phase I Gamma GLM applied to obtain the regression residuals from each observation $(t_{\ell,h+j}, w_{\ell,h+j})$. These regression residuals are denoted with $r_{\ell,h+j}$ where $j = 1, \ldots, n_\ell$ for each run $\ell = 1, \ldots, N$. This can be either Pearson, deviance or quantile residuals.

2. For an arbitrary value of α, the value of Q_1 is obtained to ensure that $100(1 - \alpha/2)$ percent of the GLM residuals lies below Q_1. Similarly, the value of Q_2 is solved to ensure $100(\alpha/2)$ percent of the simulated residuals lies below Q_2.

3. The GLM-based TBE chart is constructed with UCL $= Q_1$ and LCL $= Q_2$ as defined in (4.2). The length of inspection of each run of residuals is obtained as defined in (9.7), and denoted with LI_ℓ with $\ell = 1, \ldots, N$. The ALI_0 is calculated as the average of all computed lengths of inspection, i.e., $ALI_0 = (LI_1 + \cdots + LI_N)/N$.

4. We check if $ALI_0 = 200$. If this does not hold, we adjust the value of α and return to Step 2.

Notice that $\alpha \leq 1/200$ since $n_\ell \leq n$ for all $\ell = 1, \ldots, N$. Here, n denotes the amount of simulated observations $(y_{\ell,m+i}, x_{\ell,m+i})$ with $i = 1, \ldots, n$. The probability control limits Q_1 and Q_2 are calculated 100 times, as described in Section 9.4. The densities for Q_1 and Q_2 of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart are shown in Figures B.24, B.25 and B.26, respectively. The OC performance results of each chart in terms of ALI_1 are provided in the following section.

9.5 OC Performance results of GLM-based TBE charts

Figures 9.4, 9.5, 9.6 and 9.7 show the ALI_1 results of the Gamma GLM-based TBE charts for ZIP and ZINB scenarios 1, 2, 3 and 4, respectively. First of all, it is noticed that Pearson, deviance and quantile residuals show the exact same OC performance results. Hence, it can be concluded that the Gamma GLM-based TBE charts are not sensitive to the specific residual type that is applied, unlike the ZIP and ZINB regression-based Shewhart charts from the Chapter 7 and 8.

In addition, it is remarkable that an OC distributional shift due to increased or decreased $E[p^{OC}]$ is detected by the GLM-based TBE charts since the ALI_1 shows a decreasing trend as the distributional shift gets larger. Hence, OC scenarios with worse or improved overall process performance due to a shift in $E[p^{OC}]$ are both well detected. When comparing Figure 9.4 with Figure 9.5, it is observed that the GLM-based TBE chart has better performance with IC Scenario 1, which corresponds to ZIP distributed observation with a high proportion of zero-inflation. The same conclusion can be drawn from Figures 9.6 and 9.7, where the TBE charts perform better with IC Scenario 3, which corresponds to ZINB distributed observations with high proportion of zero-inflation.

An OC distributional shift due to increased or decreased $E[\lambda^{OC}]$ is less well detected. It is shown in Figures 9.4c and 9.6c show that the ALI_1 decreases for a OC distributional shift due to decreased $E[\lambda^{OC}]$, but an OC distributional shift due to increased $E[\lambda^{OC}]$ is not detected as fast. Hence, an OC scenario with improved process performance is detected better than an OC scenario with worse process performance in this case. In addition it is observed from Figures 9.5 and 9.7 that an OC distributional shift due to increased or decreased $E[\lambda^{OC}]$ is not detected for IC scenarios 2 and 4, which correspond to the ZIP and ZINB distributed observations with low proportion of zero inflation and high expected value. The poor performance of the Gamma GLM-based TBE chart in case of an OC distributional shift in $E[\lambda^{OC}]$ is most likely due to the TBE data transformation, in which all events $Y_i > 0$ are treated equally. The parameter λ of the ZIP and ZINB distribution represents the expected amount of detected defects, in case Y_i is not a structural zero. Hence, a small distributional shift in $E[\lambda^{OC}]$ does not affect

the amount of events $Y_i > 0$ very much, in case where the IC $E[\lambda]$ is already high. The full ALI_1 results are provided in Tables B.21, B.22 and B.24.

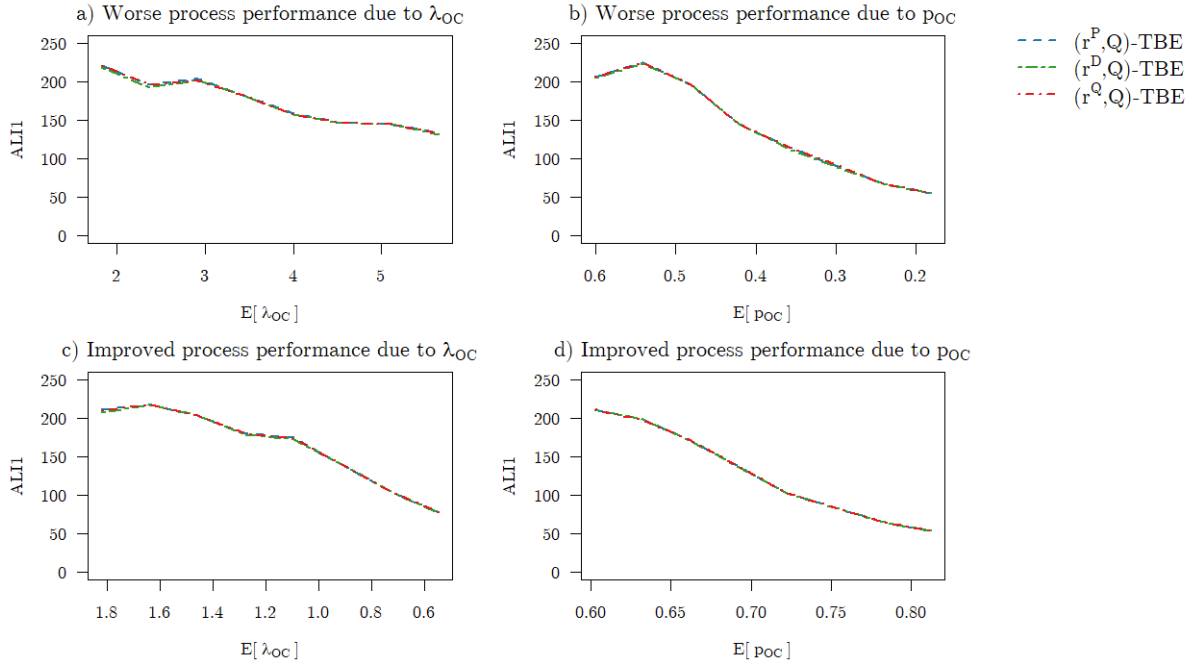

Figure 9.4: ALI_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart for IC ZIP scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

9.6 Summary

In this chapter, we introduce a Gamma GLM-based TBE chart for monitoring the time between nonconforming events, i.e., $Y_i > 0$, while correcting for the accumulated inspected weights. The distribution of TBE data is discussed in Section 9.1, where it is decided to assume TBE variable T_j follows a Gamma distribution. The Gamma GLM is defined afterwards in Section 9.2 and the monitoring procedure of TBE residuals is described in Section 9.3. The strategy for performance analysis is defined in Section 9.4 after which the ALI results are provided in Section 9.5. It is concluded based on the ALI_1 results that the Gamma GLM-based TBE charts are performing satisfactory for detecting a shift in $E[p^{OC}]$. It is also observed that the performance of the Gamma GLM-based TBE chart increases with higher proportions of zero-inflation in the IC process. Hence, this introduction of GLM-based TBE charts could provide an opportunity for monitoring high-purity processes at Dow, which inherit extreme proportions of zero-inflation. Nevertheless, it should be noted that the results in this chapter are based on the assumption that T_j follows a Gamma distribution. This is not proved for ZIP or ZINB distributed observations Y_i, such that other distributions should be considered as well when further investigating the GLM-based TBE charts.

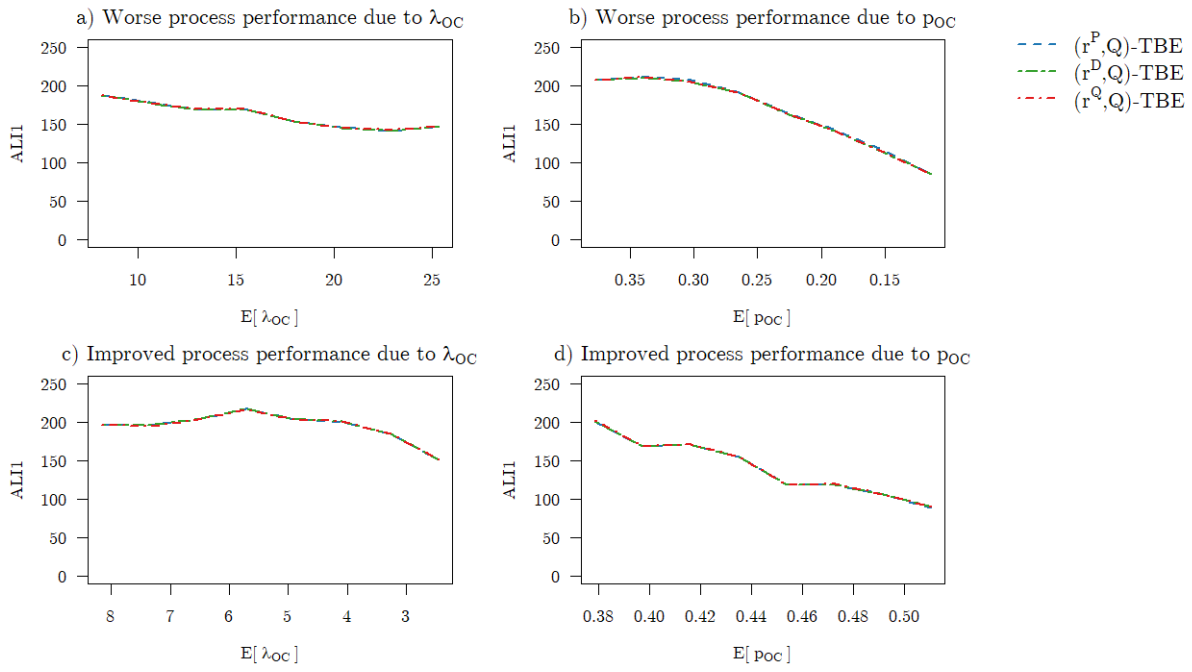

Figure 9.5: ALI_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart for IC ZIP scenario 2 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

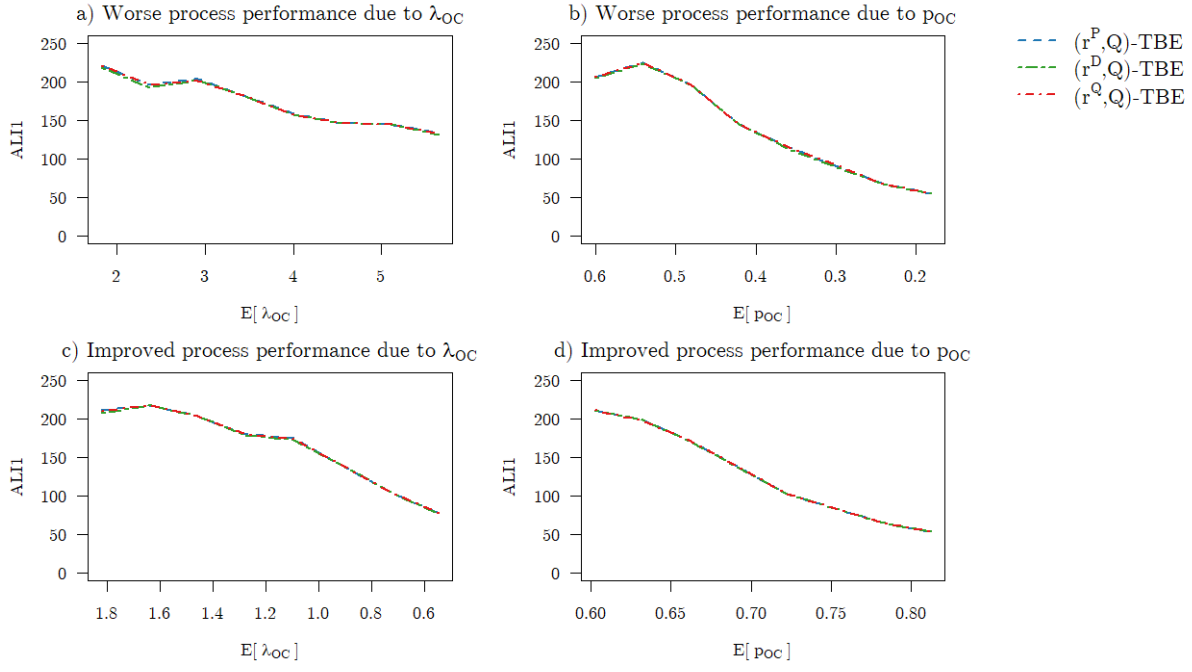

Figure 9.6: ALI_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart for IC ZINB scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

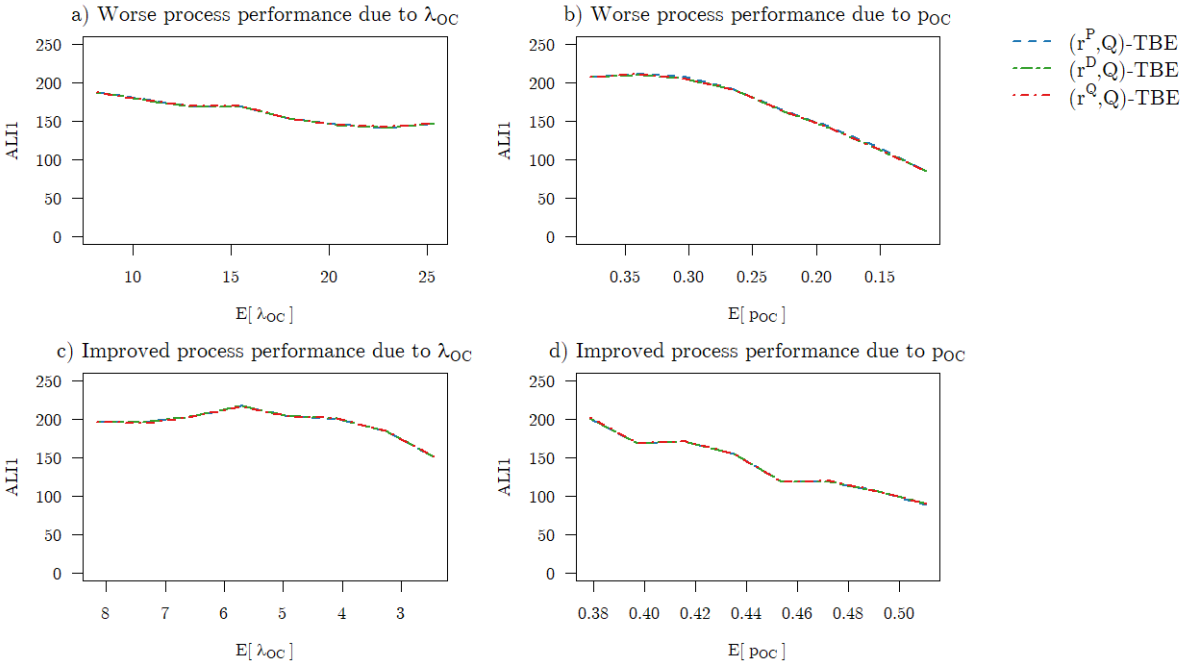

Figure 9.7: ALI_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart for IC ZINB scenario 4 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

10 | Conclusion and discussion

The main goal of this thesis is to provide a framework for identifying the most appropriate monitoring scheme for detecting contextual anomalies in zero-inflated count data. This type of data originates from monitoring high-purity processes for defects, where the amount of nonconforming observations is affected by one covariate. In order to achieve this goal, we constructed a ZIP-EWMA chart, multiple ZIP and ZINB regression-based Shewhart chart for Pearson, deviance and randomised quantile residuals, and we introduced a Gamma GLM-based TBE chart. The main conclusions from these studies are discussed in the following section, where we answer the research questions of this project.

10.1 Summary of results and conclusions

The first sub-question of this thesis is: *Based on published works in the literature, what are the established monitoring methods for detecting contextual anomalies in data that originates from monitoring high-purity processes for defects?* The answer is provided in Chapter 3. The literature study showed that regression-based control charts outperform traditional control charts when aiming to detect contextual anomalies, and that GLM-control chart outperform linear regression-based control charts when the response data follows an non-normal EDM distribution. The ZIP and ZINB regression models are employed in literature for regression-based control charts when monitoring high-purity count data. It is observed that predictive Pearson, deviance and randomised quantile residuals are used for monitoring, but a clear consensus regarding which residual type performs best in a regression-based control chart for high-purity count data does not exist. Literature on regression-based TBE charts does also not exist. Hence, the main focus of the thesis has been to provide insight in the performance of Pearson, deviance and quantile residuals in regression-based monitoring schemes for ZIP and ZINB distributed data.

The second sub-question in this research is: *How can we model the relationship between the response variable and the covariate? And what type of residuals can we use for a regression-based control chart?* The answer is provided in Chapter 6, where the one dimensional ZIP and ZINB regression models are defined with corresponding definitions for Pearson, deviance and randomised quantile residuals. It is concluded in Chapter 7 that all residual types are proper goodness of fit measures, which makes them appropriate for contextual anomaly detection. Analysis regarding the distribution of each residual type showed that randomised quantile residuals from the ZIP or ZINB regression model may be assumed to follow an independent and identical standard normal distribution. Pearson and deviance residuals show a clear non-normal distribution, such that no assumptions are made regarding their distribution throughout this project. Hence, it is concluded that ZIP and ZINB Pearson, deviance and randomised quantile residuals can all be monitored in a regression-based control chart. However, control limits can be calculated when monitoring quantile residuals while control limits must be solved numerically when monitoring Pearson or deviance residuals.

This brings us to the third sub-question, that is: *Which regression-based monitoring schemes can be used for detecting contextual anomalies in data that originates from monitoring high-purity processes for defects?* While many types of control charts exist, it is described in Section 3.4 that we focus of regression-

based Shewhart charts. Shewhart charts with both symmetric and probability control limits are defined to monitor the ZIP and ZINB regression residuals over time. In addition, a Gamma GLM-based TBE chart is introduced as a suggestion for monitoring data contains extreme proportions of zero-inflation.

The fourth sub-question is: *How can we evaluate the performance of a regression-based control chart?* The performance of each monitoring scheme is evaluated upon simulated ARL_0 and ARL_1 values, since ZIP and ZINB Pearson, deviance and quantile residuals are assumed to all follow an unknown distribution when the monitoring data becomes out-of-control. In order to evaluate how the proportion of zero-inflation affects the performance of each monitoring scheme, it is decided to consider four distinct in-control distribution and evaluate corresponding out-of-control performance. It is also described in Chapter 7 that we consider two strategies for performance evaluation for the ZIP and ZINB regression-based Shewhart charts. The baseline performance is established first, after which the OC performance of each chart is evaluated while taking into account the effects of Phase I estimation. The performance of the Gamma GLM-based TBE charts is evaluated under the assumption that all ZIP and ZINB distributional parameters are know, while we take into account that the Gamma GLM regression coefficients must be estimated.

The final sub-question of this thesis is: *Which monitoring scheme achieves the best performance when aiming to detect contextual anomalies in data that originates from monitoring high-purity processes for defects?* It is concluded from the results in Chapter 8 that the ZIP and ZINB regression-based Shewhart charts with probability control limits and deviance residuals perform best of all evaluated monitoring schemes. This is observed from both baseline performance results as well as performance results while taking into account Phase I estimation. It is concluded that the performance of the ZIP and ZINB regression-based Shewhart charts is rather sensitive to the residuals type that is monitored, and the type of control limits that are chosen, since the performance results show great differences. This is nowhere mentioned in the existing literature, but should be taken into account when designing a monitoring scheme. It is also concluded that each ZIP and ZINB regression-based Shewhart chart with Pearson, deviance and quantile residuals is detecting an out-of-control distributional shift due to increased or decreased $E[\lambda^{OC}]$, while a shift due to increased or decreased $E[p^{OC}]$ is not detected as well by any of the ZIP or ZINB regression-based Shewhart charts. It is additionally concluded that each ZIP and ZINB regression-based Shewhart chart performs better under less zero-inflated IC circumstances, even thought the differences in performance results are very modest in this thesis. The Gamma GLM-based TBE charts show satisfactory performance results in case of an out-of-control distributional shift due to increased or decreased $E[p^{OC}]$, while a shift due to increased or decreased $E[\lambda^{OC}]$ is less well detected. It is additionally concluded that the Gamma GLM-based TBE chart is not sensitive to the residual type that is monitored, since performance results are the same for each type. The Gamma GLM-based TBE chart performs better under more zero-inflated IC circumstances, such that these type of control charts are suggested for further consideration, when dealing with monitoring data contains extreme proportions of zero-inflation.

The main question of this research is: *Which monitoring scheme is most appropriate for detecting contextual anomalies in univariate count data, that originates from monitoring a specific high-purity processes?* The conclusion is the same for high-purity processes of which the monitoring data follows a ZIP or ZINB distribution with high or low proportion of zero inflation. Namely, the ZIP and ZINB regression-based Shewhart chart with deviance residuals and probability control limits performs best. A graphical representation of the conclusion is shown in Figure 10.1.

10.2 Recommendations for Dow

It has been shown this project that, the performance of a ZIP and ZINB regression-based Shewhart chart is heavily affected by the specific residual type that is monitored, and by the control limits that are

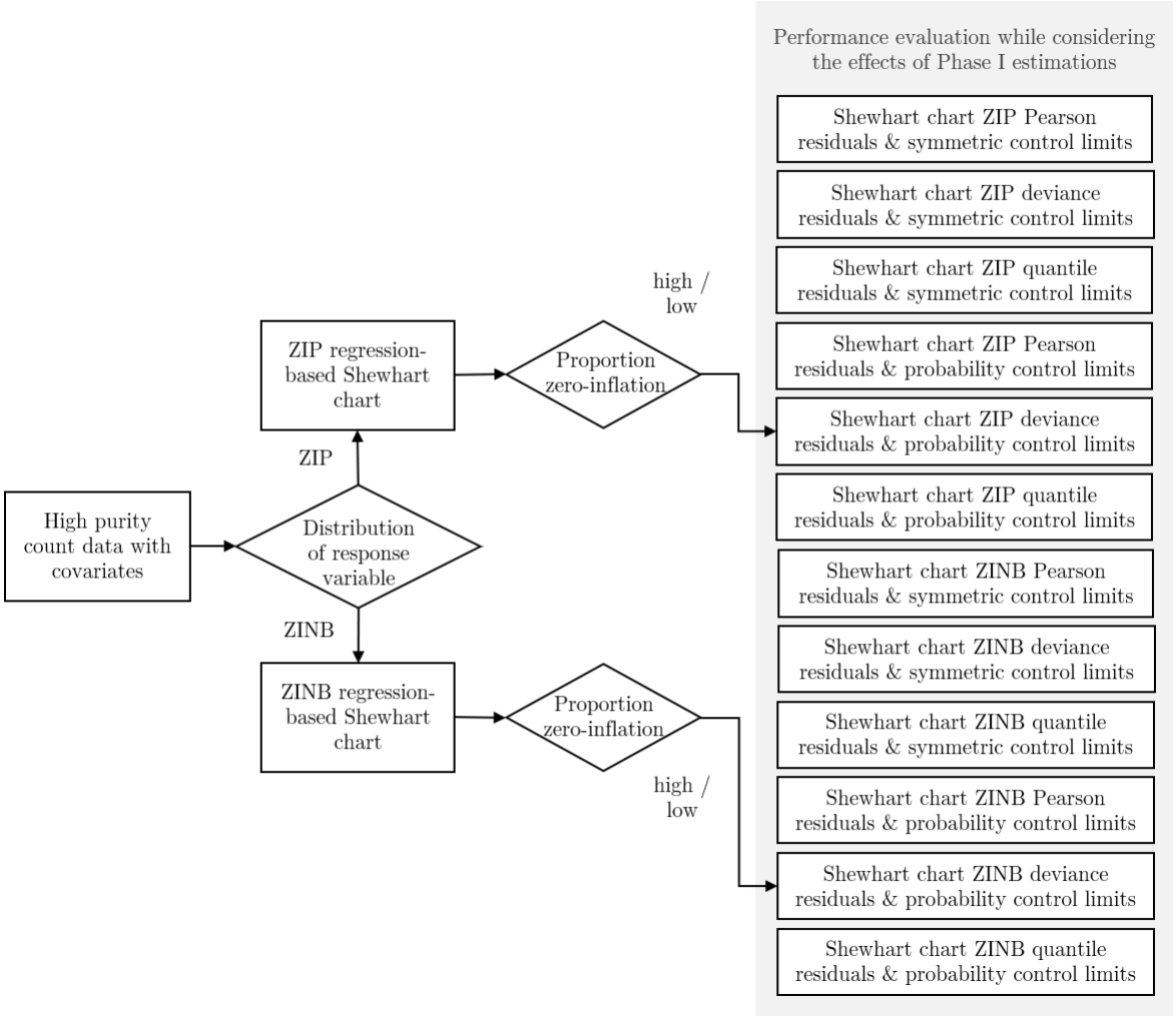

Figure 10.1: Recommended flow chart, based upon the results of this thesis.

chosen. The easiest way to construct a ZIP or ZINB regression-based Shewhart chart is by application of randomised quantile residuals, due to their standard normal distribution. Symmetric and probability control limits are the same in this case and can be fixed at a chosen value to achieve a certain ARL_0. However, the results show that the ZIP and ZINB regression-based Shewhart charts with deviance residuals and probability control limits outperform the charts for randomised quantile residuals with great difference. It is therefore recommended to Dow to implement the (r^D, Q)-Shewhart chart when monitoring zero-inflated count data that is affected by one covariate, even though solving the control limits numerically takes more effort.

In addition it is recommended to execute a comprehensive Phase I analysis before constructing any ZIP or ZINB regression-based control chart. Model selection procedures are not considered in this thesis since it is assumed that the covariate is already known. However, this is often not the case in practice and the overall performance of a regression-based control chart is affected by the goodness of fit of the Phase I model. When dealing with a large number of covariates, principal-component-analysis could be applied to transform a large number of possibly correlated covariates into a smaller number of uncorrelated covariates (see e.g. Park et al. (2018)). A model selection procedure to distinguish between Poisson, ZIP, negative binomial and ZINB distributed data is provided by Mahmood (2020), which is briefly discussed in Section 6.2. It is also recommended to ensure a large Phase I data set of preferably ≥ 1500 observations

since it is described Section 7.5.2 that the goodness of fit stabilises around 1500 observations. This is not expected to represent a limitation at Dow since high-purity processes at Dow are monitored continuously on minute bases and monitoring data is often available in abundance. Additional methods for model selection are cross validation and evaluation of randomised quantile residuals. Pearson and deviance residuals cannot be used for this purpose since their distribution is not guaranteed to be normal under perfect model fit.

Finally, it is recommended for Dow to be aware of the monitoring goal when designing a ZIP or ZINB regression-based Shewhart charts. It is observed from the results that a sudden increased amount of defect observations is well detected while an increased defect rate is hardly detected. Hence, when aiming to detect increased defect rate, it is recommended to consider an alternative control chart design or an alternative monitoring method such as the GLM-based TBE charts. Both alternatives are discussed in the following section.

10.3 Future research

It has been shown in the literature review that the application of generalized linear models in statistical process control for high-purity processes is a very current topic in the field of SPC. Therefore, there is still a considerable amount of possibilities that remain to be investigated up to now. First of all, it would be interesting revise the performance evaluation of the ZIP and ZINB regression-based Shewhart charts, while taking into account the effect of Phase I estimation, for a larger number of replications. It is discussed in Section 8.2 that a value of $N = 10,000$ would provide reliable results regarding the true effect of Phase I estimation.

Moreover, we have limited this project to the application of regression-based Shewhart charts only. However, other control charts such as the EWMA or CUSUM chart could also be employed to monitor regression residuals over time. The Shewhart chart is only considering current observations, such that small residual values due to minor cases of over- or underestimation of the regression model can never cause an OC alarm. EWMA and CUSUM charts include all historical observations in the charting statistic, such that small cases of over- or underestimation can cause an OC alarm if they remain persistent over time. It is also expected that the ZIP and ZINB regression-based EWMA or CUSUM charts are better able to detect an OC distributional shift due to parameter $E[p^{OC}]$. This because it is shown in Chapter 5 that the ZIP-EWMA chart performs satisfactory in detecting an OC shift in ZIP parameter p. The recently published paper of Mahmood et al. (2021) addresses the ZIP and ZINB regression-based EWMA and CUSUM charts for Pearson residuals. However, it would be interesting to investigate the performance deviance and randomised quantile residuals in these charts as well. Especially because the conclusions of this project state that deviance residuals outperform Pearson residuals in the ZIP and ZINB regression-based Shewhart charts.

In addition to that, it was chosen to limit the scope of this project to monitoring predictive residuals only. Monitoring predictive residuals is standard in the GLM-based SPC literature, but the existence of alternatives should be considered as well. Monitoring recursive residuals or recursive regression coefficients avoids the need for a Phase I, such that a self-starting approach could be designed. It should be noticed however that the ZIP and ZINB regression models require a large data set to obtain stable estimations for regression coefficients. This is due to the infrequent non-zero observations. Nevertheless, self-starting ZIP and ZINB regression-based control charts could present a great advantage in case Phase I data is unavailable. In addition, it is explained in Section 3.1 that residuals from a recursive linear regression model follow an independent distribution. This has not been proved so far for GLM, ZIP or ZINB regression residuals, but independence of residuals could provide an advantage when monitoring them in a Shewhart chart. Namely, independently distributed observations avoid the need to solve control limits numerically by simulation.

Then finally, the GLM-based TBE chart could be researched further as well. The design of the Gamma GLM-based TBE chart in this thesis is completely based on the assumption that the time-between-events of ZIP or ZINB distributed data is Gamma distributed. This is however not necessarily true in practice, and the distribution of the TBE data depends on the distribution of defect observations. Nevertheless, we obtain promising performance results from the Gamma GLM-based TBE charts, which seem to improve with higher proportions of zero-inflation in the IC data. Hence, GLM-based TBE charts could present an opportunity for monitoring very high-purity processes which include one or more covariates. Currently, there is no literature available regrading regression-based TBE charts. This is most likely due to the fact that not all covariates can be aggregated as easily as in the context of plastic pellet production. Still, regression-based TBE charts would be interesting to further explore. For example, this thesis is limited to Gamma GLM-based TBE Shewhart charts with aggregation level $r = 1$. However, other aggregation levels should be evaluated as well before designing a GLM-based TBE chart in practice. In addition, GLM-based TBE EWMA or CUSUM charts could also be designed and other EDM distributions can be considered for the TBE data as well.

Bibliography

W. Albers and W.C.M. Kallenberg. Are estimated control charts in control? *Statistics*, 38(1):67–79, 2004.

V. Alevizakos and C. Koukouvinos. Monitoring of zero-inflated Poisson processes with EWMA and DEWMA control charts. *Quality and Reliability Engineering International*, 36(1):88–111, 2020.

R.L. Brown, J. Durbin, and J.M. Evans. Techniques for testing the constancy of regression relationships over time. *Journal of the Royal Statistical Society. Series B (Methodological)*, 37(2):149–192, 1975.

T. Calvin. Quality control techniques for 'zero defects'. *IEEE Transactions on Components, Hybrids, and Manufacturing Technology*, 6(3):323–328, 1983.

S. Chakraborti. Run length distribution and percentiles: the Shewhart chart with unknown parameters. *Quality Engineering*, 19(2):119–127, 2007.

C.J. Chu, M. Stinchcombe, and H. White. Monitoring structural change. *Econometrica: Journal of the Econometric Society*, 64(5):1045–1065, 1996.

J.M. Dufour. Recursive stability analysis of linear regression relationships. *Journal of Econometrics*, 19 (1):31–76, 1982.

P.K. Dunn and G.K. Smyth. Randomized quantile residuals. *Journal of Computational and Graphical Statistics*, 5(3):236–244, 1996.

P.K. Dunn and G.K. Smyth. *Generalized Linear Models with examples in R.* Springer, 2018.

W. Feller. On the normal approximation to the binomial distribution. *Ann. Math. Stat.*, 16:319–329, 1945.

C. Feng, L. Li, and A. Sadeghpour. A comparison of residual diagnosis tools for diagnosing regression models for count data. *BMC Medical Research Methodology*, 20(1):1–21, 2020.

T. N. Goh. A control chart for very high yield processes. *Quality Assurance*, 13(1):18–22, 1987.

D.M. Hawkins, P. Qiu, and C.W. Kang. The changepoint model for statistical process control. *Journal of Quality Technology*, 35(4):355–366, 2003.

D.C. Heilbron. Zero-altered and other regression models for count data with added zeros. *Biometrical Journal*, 36(5):531–547, 1994.

A. Jamal, T. Mahmood, M. Riaz, and H.M. Al-Ahmadi. GLM-based flexible monitoring methods: An application to real-time highway safety surveillance. *Symmetry*, 13(2):362, 2021.

D. Jearkpaporn, D.C. Montgomery, G.C. Runger, and C.M. Borror. Process monitoring for correlated gamma-distributed data using generalized-linear-model-based control charts. *Quality and Reliability Engineering International*, 19(6):477–491, 2003.

I.T. Jolliffe. Sample sizes and the central limit theorem: The Poisson distribution as an illustration. *Amer. Statistician*, 49(3):269–269, 1995.

B. Jørgensen. Proper dispersion models. *Brazilian Journal of Probability and Statistic*, 11(2):89–128, 1997.

J.M. Juran. Early SQC: A historical supplement. *Quality Progress*, 30(9):73–82, 1997.

S. Knoth. *spc: Statistical Process Control*, 2021. R package version 0.6.5 – Calculation of ARL and Other Control Chart Performance Measures.

L. Koralov and Y.G. Sinai. *Theory of Probability and Random Processes*. Springer Science & Business Media, 2007.

D. Lambert. Zero-inflated Poisson regression, with an application to defects in manufacturing. *Technometrics*, 34(1):1–14, 1992.

T. Mahmood. Generalized linear model based monitoring methods for high-yield processes. *Quality and Reliability Engineering International*, 36(5):1570–1591, 2020.

T. Mahmood, N. Balakrishnan, and M. Xie. The generalized linear model-based exponentially weighted moving average and cumulative sum charts for the monitoring of high-quality processes. *Applied Stochastic Models in Business and Industry*, pages 1–22, 2021.

B.J. Mandel. The regression control chart. *Journal of Quality Technology*, 1(1):1–9, 1969.

P. McCullagh and J.A. Nelder. *Generalized linear models*. Routledge, 2019.

F.M. Megahed and L.A. Jones-Farmer. Statistical perspectives on "Big Data". In *Frontiers in statistical quality control 11*, pages 29–47. Springer, 2015.

J.A. Nelder and R.W.M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, 135(3):370–384, 1972.

E.S. Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

K. Park, J. Kim, and D. Jung. GLM-based statistical control r-charts for dispersed count data with multicollinearity between input variables. *Quality and Reliability Engineering International*, 34(6): 1103–1109, 2018.

K. Park, D. Jung, and J. Kim. Control charts based on randomized quantile residuals. *Applied Stochastic Models in Business and Industry*, 36(4):716–729, 2020.

P. Qiu. *Introduction to Statistical Process Control*. CRC Press: Broca raton, 2013.

P. Qiu. Big data? Statistical process control can help! *The American Statistician*, 74(4):329–344, 2020.

C. Rizzo, S. Chin, E. van den Heuvel, and A. Di Bucchianico. Performance measures of discrete and continuous time-between-events control charts. *Quality and Reliability Engineering International*, 36 (8):2754–2768, 2020.

S.W. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3):239–250, 1959.

J.R. Schaffer and M. Kim. Number of replications required in control chart Monte Carlo simulation studies. *Communications in Statistics—Simulation and Computation®*, 36(5):1075–1087, 2007.

W.A. Shewhart. The application of statistics as an aid in maintaining quality of a manufactured product. *Journal of the American Statistical Association*, 20(152):546–548, 1925.

A.N. Shiryaev. On optimum methods in quickest detection problems. *Theor. Prob. Appl.*, 8(1):22–46, 1963.

L. Shu, F. Tsung, and K. Tsui. Effects of estimation errors on cause-selecting charts. *IIE transactions*, 37(6):559–567, 2005.

K.R. Skinner, D.C. Montgomery, and G.C. Runger. Process monitoring for multiple count data using generalized linear model-based control charts. *International Journal of Production Research*, 41(6): 1167–1180, 2003.

K.R. Skinner, D.C. Montgomery, and G.C. Runger. Generalized linear model-based control charts for discrete semiconductor process data. *Quality and Reliability Engineering International*, 20(8):777–786, 2004.

J. Stauffer. SQC Before Deming: The works of Walter Shewhart. *Journal of Applied Management and Entrepreneurship*, 8(4):86, 2003.

O. Van Dalen. Statistical monitoring of wind turbines. Bachelor thesis, Eindhoven University of Technology, Department of Mathematics and Computer Science, Eindhoven, The Netherlands, 2018.

Q.H. Vuong. Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society*, 57(2):307–333, 1989.

M. Weese, W. Martinez, F.M. Megahed, and L.A. Jones-Farmer. Statistical learning methods applied to process monitoring: An overview and perspective. *Journal of Quality Technology*, 48(1):4–24, 2016.

M. Xie, B. He, and T.N. Goh. Zero-inflated Poisson model in statistical process control. *Computational statistics & Data Analysis*, 38(2):191–201, 2001.

M. Xie, T.N. Goh, and V. Kuralmani. *Statistical models and control charts for high-quality processes*. Springer Science & Business Media, 2002a.

M. Xie, T.N. Goh, and P. Ranjan. Some effective control chart procedures for reliability monitoring. *Reliability Engineering & System Safety*, 77(2):143–150, 2002b.

A. Zeileis, F. Leisch, C. Kleiber, and K. Hornik. Monitoring structural change in dynamic econometric models. *Journal of Applied Econometrics*, 20(1):99–121, 2005.

A | Appendix: Lemmas and proofs

This appendix contains the mathematical details that support the thesis. A proof of the ZIP, ZINB and Gamma expected value and variance is provided in Section A.1, and a proof of the normality of quantile residuals is provided in Section A.2.

A.1 ZIP, ZINB and Gamma expected value and variance

Lemma A.1.1 (**Expected value of the zero-inflated Poisson distribution**). *Let us assume random variable Y follows a zero-inflated Poisson distribution with probability mass function*

$$P(Y = y) = \begin{cases} p + (1-p)e^{-\lambda} & \text{if } y = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y > 0 \end{cases}$$

where both λ and p are known. Then the expected value of Y is defined as $E[Y] = (1-p)\lambda$.

Proof. The expected value of Y is defined as:

$$\begin{aligned} \mathrm{E}[Y] &= \sum_{y=0}^{\infty} y \cdot P(Y = y) \\ &= 0 \cdot \left[p + (1-p)e^{-\lambda}\right] + (1-p)\sum_{y=1}^{\infty} y\frac{e^{-\lambda}\lambda^y}{y!} \\ &= (1-p)\lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \\ &= (1-p)\lambda \end{aligned}$$

where the infinite sum equals the Taylor expansion of the exponential function e^{λ}. Hence, $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda}$ which ensures $\mathrm{E}[Y] = (1-p)\lambda$. This concludes the proof. □

Lemma A.1.2 (**Variance of the zero-inflated Poisson distribution**). *Let us assume random variable Y follows a zero-inflated Poisson distribution with probability mass function*

$$P(Y = y) = \begin{cases} p + (1-p)e^{-\lambda} & \text{if } y = 0 \\ (1-p)\frac{e^{-\lambda}\lambda^y}{y!} & \text{if } y > 0 \end{cases}$$

where both λ and p are known. Then the variance of Y is defined as$\mathrm{Var}(Y) = (1-p)(\lambda + p\lambda^2)$.

Proof. The variance of a random variable is defined by $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$. We have from Lemma

A.1.1 that $E[Y] = (1-p)\lambda$. We proceed with the definition of $E[Y^2]$ as follows.

$$\begin{aligned} E[Y^2] &= \sum_{y=0}^{\infty} y^2 \cdot P(Y=y) \\ &= (1-p)\sum_{y=1}^{\infty} y^2 \frac{e^{-\lambda}\lambda^y}{y!} \\ &= (1-p)(\lambda+\lambda^2)e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} \end{aligned}$$

where we substitute $y = k+1$ in the second step and the infinite sum equals the Taylor expansion of the exponential function e^{λ}. Hence, $\sum_{k=0}^{\infty}\frac{\lambda^k}{k!} = e^{\lambda}$ which ensures $\mathrm{E}[Y^2] = (1-p)(\lambda+\lambda^2)$. We can use this definition, and the result from A.1.1 to obtain a definition for the variance.

$$\begin{aligned} \mathrm{Var}(Y) &= E[Y^2] - (E[Y])^2 \\ &= (1-p)(\lambda+\lambda^2) - (1-p)^2\lambda^2 \\ &= (1-p)(\lambda+p\lambda^2) \end{aligned}$$

This concludes the proof. □

Lemma A.1.3 (**Expected value of the zero-inflated negative binomial distribution**). *Let us assume random variable Y follows a zero-inflated negative binomial distribution with probability mass function*

$$P(Y=y) = \begin{cases} p + (1-p)\left(1+\frac{\lambda}{\tau}\right)^{-\tau} & \text{if } y=0 \\ (1-p)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}\left(1+\frac{\lambda}{\tau}\right)^{-\tau}\left(1+\frac{\tau}{\lambda}\right)^{-y} & \text{if } y>0 \end{cases}$$

where parameters τ, λ and p are known. Then the expected value of Y is defined as $E[Y] = (1-p)\lambda$.

Proof. We approach this proof by using the definition of the binomial coefficient, $\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)} = \binom{y+\tau-1}{y} = \frac{(y+\tau-1)!}{y!(\tau-1)!}$, and defining the expected value of Y as:

$$\begin{aligned} E[Y] &= \sum_{y=0}^{\infty} y \cdot P(Y=y) \\ &= \sum_{y=1}^{\infty}(1-p)\frac{\Gamma(y+\tau)}{(y-1)!\Gamma(\tau)}\left(1+\frac{\lambda}{\tau}\right)^{-\tau}\left(1+\frac{\tau}{\lambda}\right)^{-y} \\ &= (1-p)\tau\sum_{y=1}^{\infty}\binom{y+\tau-1}{y-1}\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)^{y} \end{aligned}$$

since $y\binom{y+\tau-1}{y} = \tau\binom{y+\tau-1}{y-1}$ by definition of the binomial coefficient. We proceed by rewriting the definition of $E[Y]$ into the exact form that is needed to apply the binomial theorem. This requires substituting

$m = y - 1$. The binomial theorem is finally applied in the one but last step of the following definition.

$$\begin{aligned}
E[Y] &= (1-p)\tau \sum_{m=0}^{\infty} \binom{m+\tau}{m} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{m+1} \\
&= (1-p)\frac{\lambda}{\tau+\lambda}\tau \sum_{m=0}^{\infty} \binom{m+\tau}{m} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{m} \\
&= (1-p)\lambda \sum_{m=0}^{\infty} \binom{m+\tau}{m} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau+1} \left(\frac{\lambda}{\tau+\lambda}\right)^{m} \\
&= (1-p)\lambda \left(\frac{\tau}{\tau+\lambda} + \frac{\lambda}{\tau+\lambda}\right)^{m+\tau} \\
&= (1-p)\lambda
\end{aligned}$$

By use of the binomial theorem we have found that $E[Y] = (1-p)\lambda$. This concludes the proof. □

Lemma A.1.4 (**Variance of the zero-inflated negative binomial distribution**). *Let us assume random variable Y follows a zero-inflated negative binomial distribution with probability mass function*

$$P(Y = y) = \begin{cases} p + (1-p)\left(1 + \frac{\lambda}{\tau}\right)^{-\tau} & \text{if } y = 0 \\ (1-p)\frac{\Gamma(y+\tau)}{y!\Gamma(\tau)}\left(1 + \frac{\lambda}{\tau}\right)^{-\tau}\left(1 + \frac{\tau}{\lambda}\right)^{-y} & \text{if } y > 0 \end{cases}$$

where parameters τ, λ and p are known. Then the variance of Y is defined as $\mathrm{Var}(Y) = \lambda(1-p)(1 + p\lambda + \lambda/\tau)$.

Proof. The variance of a random variable is defined by $\mathrm{Var}(Y) = E[Y^2] - E[Y]^2$. We have from Lemma A.1.3 that $E[Y] = (1-p)\lambda$. We proceed with the definition of $E[Y^2]$ as follows.

$$\begin{aligned}
E[Y^2] &= \sum_{y=0}^{\infty} y^2 \cdot P(Y = y) \\
&= \sum_{y=1}^{\infty} (1-p) y^2 \frac{\Gamma(y+\tau)}{(y)!\Gamma(\tau)} \left(1 + \frac{\lambda}{\tau}\right)^{-\tau} \left(1 + \frac{\tau}{\lambda}\right)^{-y} \\
&= (1-p)\tau \sum_{y=1}^{\infty} y \binom{y+\tau-1}{y-1} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{y} \\
&= (1-p)\tau \sum_{m=0}^{\infty} (m+1) \binom{m+\tau}{m} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{m+1}
\end{aligned}$$

We can rewrite the binomial coefficient in this definition as:

$$\begin{aligned}
\binom{m+\tau}{m}(m+1) &= \frac{(m+\tau)!}{m!\tau!}(m+1) = \frac{(m+\tau)!}{m!\tau!}m + \frac{(m+\tau)!}{m!\tau!} = \frac{(m+\tau)!}{(m-1)!\tau!} + \frac{(m+\tau)!}{m!\tau!} \\
&= \frac{(m+\tau)!}{(m-1)!(\tau+1)!}(\tau+1) + \frac{(m+\tau)!}{m!\tau!} = \binom{m+\tau}{m-1}(\tau+1) + \binom{m+\tau}{m}
\end{aligned}$$

Hence, the definition of $E[Y^2]$ can be written as:

$$\begin{aligned}
E[Y^2] = (1-p)\tau(\tau+1) &\sum_{m=0}^{\infty} \binom{m+\tau}{m-1} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{m+1} \\
+ (1-p)\tau &\sum_{m=0}^{\infty} \binom{m+\tau}{m} \left(\frac{\tau}{\tau+\lambda}\right)^{\tau} \left(\frac{\lambda}{\tau+\lambda}\right)^{m+1}
\end{aligned}$$

By the binomial identity we have that $\binom{-n}{k} = \binom{n+k-1}{k}(-1)^k$. Therefore,

$$\begin{aligned}
E[Y^2] &= (1-p)\tau(\tau+1)\sum_{m=0}^{\infty}\binom{-(\tau+2)}{m-1}(-1)^{m-1}\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)^{m+1} \\
&\quad + (1-p)\tau\sum_{m=0}^{\infty}\binom{-(\tau+1)}{m}(-1)^{m}\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)^{m+1} \\
&= (1-p)\tau(\tau+1)\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)^{2}\sum_{m=0}^{\infty}\binom{-(\tau+2)}{m-1}\left(\frac{-\lambda}{\tau+\lambda}\right)^{m-1} \\
&\quad + (1-p)\tau\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)\sum_{m=0}^{\infty}\binom{-(\tau+1)}{m}\left(\frac{-\lambda}{\tau+\lambda}\right)^{m}.
\end{aligned}$$

By the binomial theorem, we have that:

$$\begin{aligned}
E[Y^2] &= (1-p)\tau(\tau+1)\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)^{2}\left(1-\frac{\lambda}{\tau+\lambda}\right)^{-(\tau+2)} \\
&\quad + (1-p)\tau\left(\frac{\tau}{\tau+\lambda}\right)^{\tau}\left(\frac{\lambda}{\tau+\lambda}\right)\left(1-\frac{\lambda}{\tau+\lambda}\right)^{-(\tau+1)} \\
&= (1-p)\tau(\tau+1)\left(\frac{\lambda}{\tau+\lambda}\right)^{2}\left(\frac{\tau}{\tau+\lambda}\right)^{-2} + (1-p)\tau\left(\frac{\lambda}{\tau+\lambda}\right)\left(\frac{\tau}{\tau+\lambda}\right)^{-1} \\
&= (1-p)\lambda\left[\lambda+\lambda/\tau+1\right]
\end{aligned}$$

Now that we have a definition for $E[Y^2]$, we can define the variance by

$$\begin{aligned}
\mathrm{Var}(Y) &= E[Y^2] - (E[Y])^2 \\
&= (1-p)\lambda\left[\lambda+\lambda/\tau+1\right] - (1-p)^2\lambda^2 \\
&= (1-p)\lambda\left[1+p\lambda+\lambda/\tau\right].
\end{aligned}$$

This concludes the proof. □

Lemma A.1.5 (**Expected value of the Gamma distribution**). *Let us assume that random variable T follows a Gamma distribution with shape parameter k and scale parameter s. The probability distribution function at point t is defined as*

$$f_T(t,k,s) = \frac{1}{s^k\Gamma(k)}t^{k-1}e^{-t/s}.$$

Then the variance of T equals $\mathrm{Var}[T] = ks^2$.

Proof. The expected value of T is defined as

$$\mathrm{E}\left[T\right] = \int_0^{\infty} t f_T(t,k,s)dt = \frac{1}{s^k\Gamma(k)}\int_0^{\infty} t^k e^{-t/s}dt$$

Substituting $z = t/s$ leads to

$$= \frac{s}{\Gamma(k)}\int_0^{\infty} z^k e^{-z}dz = \frac{s\Gamma(k+1)}{\Gamma(k)} = \frac{sk\Gamma(k)}{\Gamma(k)} = ks$$

where in the second step we use the definition of the Gamma function $\Gamma(k) = \int_0^{\infty} z^{k-1}e^{-z}dz$, and in the third we use the Gamma difference equation $\Gamma(k+1) = \Gamma(k)$, which follows directly from the Gamma function. □

Lemma A.1.6 (**Variance of the Gamma distribution**). *Let us assume that random variable T follows a gamma distribution with shape parameter k and scale parameter s. The probability distribution function at point t is defined as*

$$f_T(t,k,s) = \frac{1}{s^k\Gamma(k)}t^{k-1}e^{-t/s}.$$

Then the expected value of T equals $E[T] = ks$.

Proof. The variance of a random variable is defined by $\mathrm{Var}(T) = E[T^2] - E[T]^2$. By Lemma A.1.5 we have that $E[T]^2 = (ks)^2$. The value of $E[T^2]$ is defined as

$$\mathrm{E}\left[T^2\right] = \int_0^\infty t^2 f_T(t,k,s)dt = \frac{1}{s^k\Gamma(k)}\int_0^\infty t^{k+1}e^{-t/s}dt$$

Substituting $z = t/s$ leads to

$$= \frac{s^2}{\Gamma(k)}\int_0^\infty z^{k+1}e^{-z}dz = \frac{s^2\Gamma(k+2)}{\Gamma(k)} = \frac{s^2k(k+1)\Gamma(k)}{\Gamma(k)} = k(k+1)s^2$$

where in the second step we use the definition of the Gamma function $\Gamma(k) = \int_0^\infty z^{k-1}e^{-z}dz$, and in the third we use the Gamma difference equation $\Gamma(k+2) = k(k+1)\Gamma(k)$, which follows directly from the Gamma function. Hence, the variance is defined as

$$\mathrm{Var}(T) = E[T^2] - E[T]^2 = k(k+1)s^2 - k^2s^2 = ks^2$$

This concludes the proof. □

A.2 Normality of quantile residuals

Lemma A.2.1 (**Normality of quantile residuals, continuous case**). *Let Y be a continuous random variable with $F_Y(y) = P(Y \leq y)$. Let $r_Q = \Phi^{-1}(F_Y(Y))$ where $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. Then r_Q follows a standard normal distribution.*

Proof. Let $U = F_Y(Y)$ and $F_Y^{-1}(u) \equiv \inf\{u : F_Y(y) \geq u\}$ for $u \in (0,1)$. Then:

$$\begin{aligned} F_U(u) &= P(U \leq u) \\ &= P(F_Y(Y) \leq u) \\ &= P(Y \leq F_Y^{-1}(u)) \\ &= F_Y(F_Y^{-1}(u)) \\ &= u \end{aligned}$$

Hence, we have that U follows a [0,1] uniform distribution. By definition of U, we have that $r_Q = \Phi^{-1}(U)$. Thus:

$$\begin{aligned} F_{r_Q}(r) &= P(r_Q \leq r) \\ &= P(\Phi^{-1}(U) \leq r) \\ &= P(U \leq \Phi(r)) \\ &= \Phi(r) \end{aligned}$$

Hence, r_Q follows a standard normal distribution. This concludes the proof. □

Lemma A.2.2 (**Normality of randomised quantile residuals, discrete case**). *Let Y be a discrete random variable with $F_Y(y) = P(Y \leq y)$. Let $b = F_Y(y)$ and $a = \lim_{\varepsilon\to 0^-} F_Y(y+\varepsilon)$. Hence, $a, b \in (0,1]$*

and $a < b$. Let $r_Q = \Phi^{-1}(U)$ where $U \sim$ Uniform(a, b), and $\Phi(\cdot)$ represents the cumulative distribution function of the standard normal distribution. Then r_Q follows a standard normal distribution.

Proof. We have:

$$\begin{aligned} F_{r_Q}(r) &= P(r_Q \leq r) \\ &= P(\Phi^{-1}(U) \leq r) \\ &= P(U \leq \Phi(r)) \end{aligned}$$

Let us denote S as the support of Y. Then we can condition on the value of Y, to obtain:

$$\begin{aligned} F_{r_Q}(r) &= \sum_{y \in S} P(U \leq \Phi(r) | Y = y) P(Y = y) \\ &= \sum_{y \in S} P(U \leq \Phi(r) | \lim_{\varepsilon \to 0^-} F_Y(y + \varepsilon) < U \leq F_Y(y)) P(Y = y) \\ &= \sum_{y \in S} P(U \leq \Phi(r) | a_y < U \leq b_y) P(Y = y) \end{aligned}$$

where $a_y = \lim_{\varepsilon \to 0^-} F_Y(y + \varepsilon)$ and $b_y = F_Y(y)$. For a given value $\Phi(r)$, let us define the set $S_1 \subseteq S$ such that $\{y \in S1 : \Phi(r) < a_y\}$. Notice that for discrete distributions, there is a unique value $\tilde{y} \in S$ for which $a_{\tilde{y}} \leq \Phi(r) \leq b_{\tilde{y}}\}$. Finally, let us define the subset $S_2 \subseteq S$ such that $\{y \in S_2 : b_y < \Phi(r)\}$. Notice that $\{S_1 \cup \tilde{y} \cup S_2\} = S$. Also notice that and $y_1 > \tilde{y}$ and $y_2 < \tilde{y}$ for all $y_1 \in S_1$ and $y_2 \in S_2$. Then we have:

$$\begin{aligned} F_{r_Q}(r) &= \sum_{y \in S_1} P(U \leq \Phi(r) \,|\, \Phi(r) < a_y) \cdot P(Y = y) + P(U \leq \Phi(r) \,|\, a_{\tilde{y}} \leq \Phi(r) < b_{\tilde{y}}) \cdot P(Y = \tilde{y}) \\ &\quad + \sum_{y \in S_2} P(U \leq \Phi(r) \,|\, \Phi(r) \geq b_y) \cdot P(Y = y) \\ &= \sum_{y \in S_1} 0 \cdot P(Y = y) + \frac{\Phi(r) - a_{\tilde{y}}}{b_{\tilde{y}} - a_{\tilde{y}}} P(Y = \tilde{y}) + \sum_{y \in S_2} 1 \cdot P(Y = y) \\ &= \frac{\Phi(r) - P(Y < \tilde{y})}{P(Y = \tilde{y})} P(Y = \tilde{y}) + P(Y < \tilde{y}) \\ &= \Phi(r) \end{aligned}$$

since $b_{\tilde{y}} - a_{\tilde{y}} = \lim_{\varepsilon \to 0^-} F_Y(\tilde{y}) - F_Y(\tilde{y} + \varepsilon) = P(Y = \tilde{y})$ and $a_{\tilde{y}} = \lim_{\varepsilon \to 0^-} F_Y(\tilde{y} + \varepsilon) = P(Y < \tilde{y})$. Hence, we have ultimately found that r_Q follows a standard normal distribution. This concludes the proof. □

B | Additional results

This appendix contains additional results that were used in the project but omitted from the main report. Section B.1 contains additional results of the ZIP-EWMA chart while Section B.2 shows the analysis results of regarding distribution of ZIP and ZINB Pearson, deviance and randomised quantile residuals in IC scenarios 2, 3 and 4. Sections B.3 and B.4 show the baseline performance and performance while estimating Phase I effects results for IC scenarios 3 and 4. The corresponding ARL and SDRL values are provided in Section B.5 after which additional results of the Gamma GLM-based TBE charts are provided in Section B.6.

B.1 Additional results of ZIP-EWMA

p	**Predefined ARL**	L	n	**Obtained ARL**	**SDRL**	**Nr. runs with no OC signal**	**% of total**
0.3	200	2.5771	400	174.1836	132.6692	1351	13.51
			1000	202.5332	194.4338	61	0.61
			2000	197.8801	198.3860	1	0.01
	370	2.8360	740	316.2641	243.8205	1307	13.07
			1850	365.5290	358.0128	68	0.68
			3700	369.8448	363.8795	1	0.01
	500	2.9547	1000	428.3289	327.9250	1273	12.73
			2500	482.4411	461.5741	51	0.51
			5000	491.0324	489.8747	0	0.0
0.5	200	2.7115	400	175.4360	132.6517	1424	14.24
			1000	207.8926	199.5865	77	0.77
			2000	205.4243	203.5562	0	0.00
	370	3.0107	740	319.7327	247.3616	1414	14.14
			1850	365.8661	360.2878	79	0.79
			3700	369.8831	367.7087	0	0.00
	500	3.1554	1000	432.5587	332.3471	1385	13.85
			2500	505.3993	488.2312	79	0.79
			5000	508.2210	501.1221	0	0.00
0.8	200	3.1804	400	170.0864	133.4355	1375	13.75
			1000	199.5170	195.9570	74	0.74
			2000	196.7787	198.6387	0	0.00
	370	3.5951	740	315.6735	244.3463	1311	13.11
			1850	357.9643	350.2038	58	0.58
			3700	354.6419	361.5027	0	0.00
	500	3.8186	1000	427.6744	331.4937	1349	13.49
			2500	491.2253	475.2914	70	0.70
			5000	504.2220	508.9362	1	0.01

Table B.1: Results of ARL computations with the total number of observations n equal to 2, 5 and 10 times the predefined ARL value.

B.2 Distribution of residuals in IC scenarios 2,3 and 4

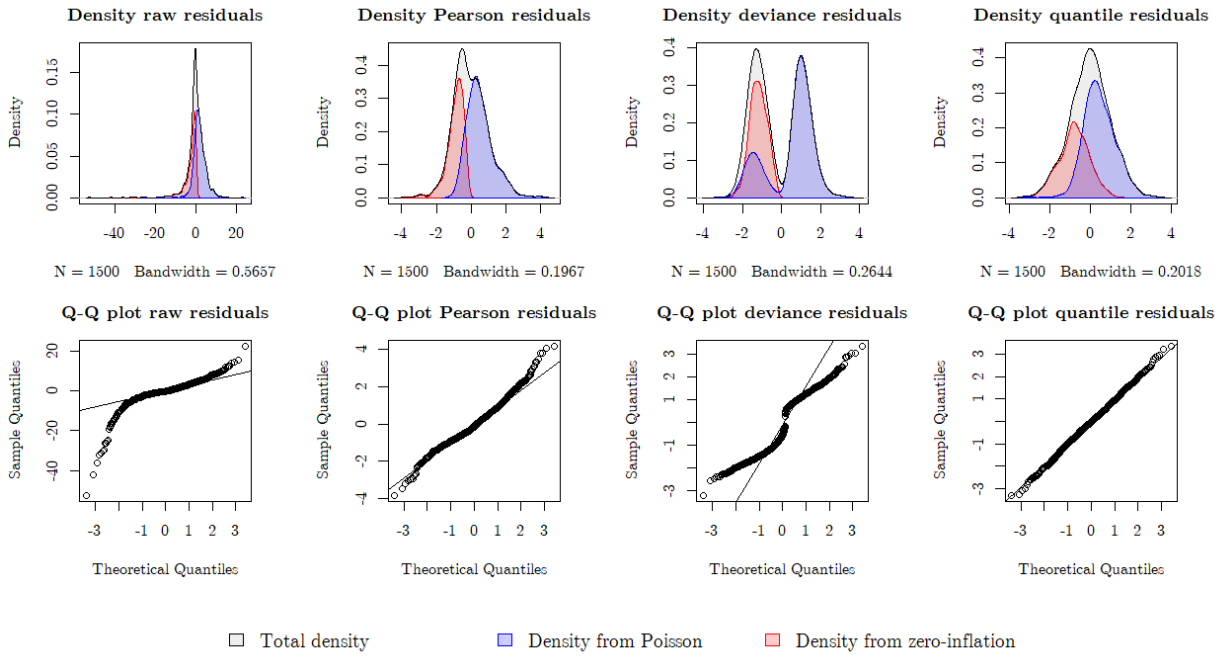

Figure B.1: Density and Q-Q plots of ZIP regression residuals for IC Scenario 2 (ZIP), indicating whether the residuals originates from the zero-inflation (red) or the Poisson distribution (blue).

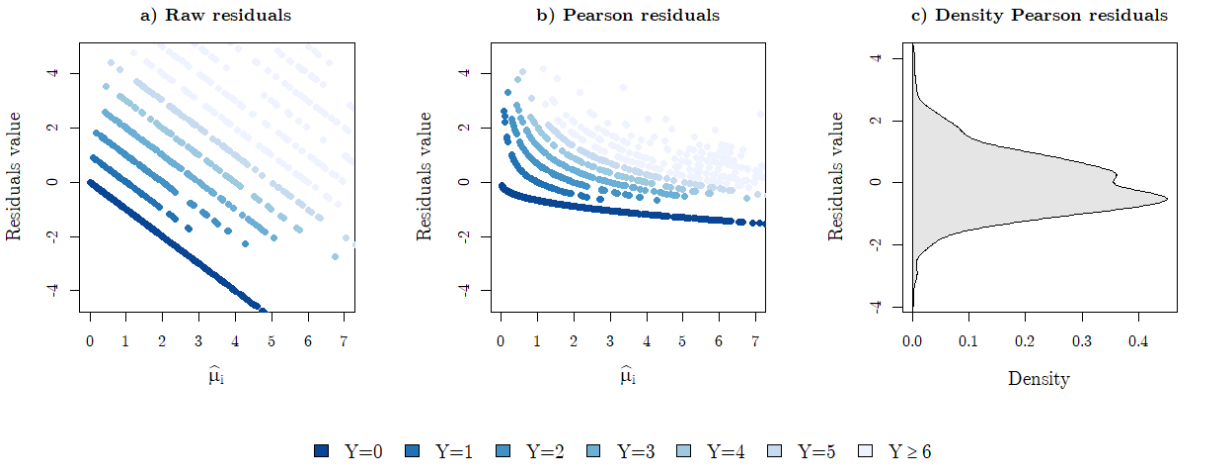

Figure B.2: Breakdown of Pearson residuals in IC Scenario 2 (ZIP), with: a) Raw residuals $r_j = y_j - \hat{\mu}_j$ plotted against prediction $\hat{\mu}_j$, b) Pearson residuals r_j^P plotted against prediction $\hat{\mu}_j$ and c) a density plot of Pearson residuals.

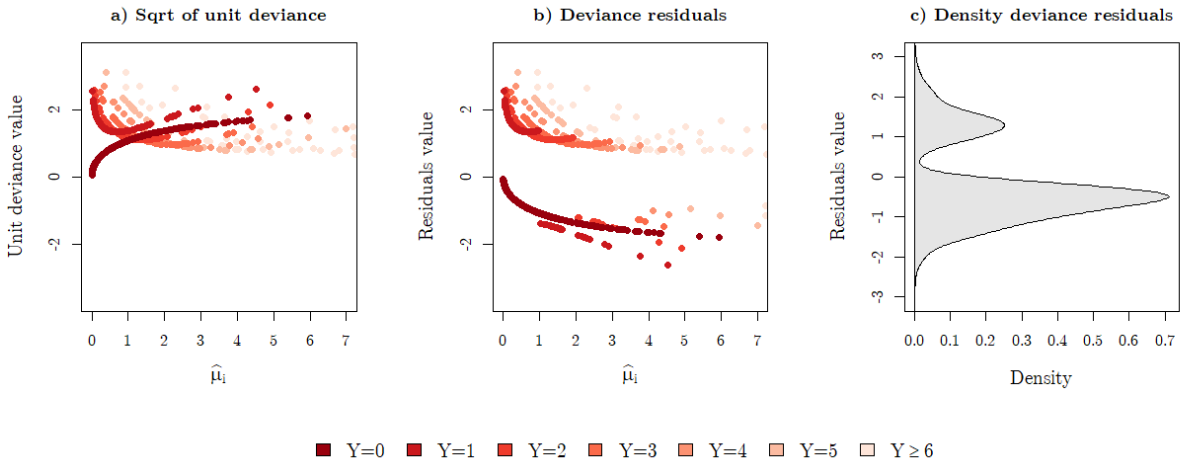

Figure B.3: Breakdown of deviance residuals in IC Scenario 2 (ZIP), with: a) the squared unit deviance $\sqrt{d_j}$ plotted against prediction $\hat{\mu}_j$, b) deviance residuals r_j^D plotted against prediction $\hat{\mu}_j$ and c) a density plot of deviance residuals.

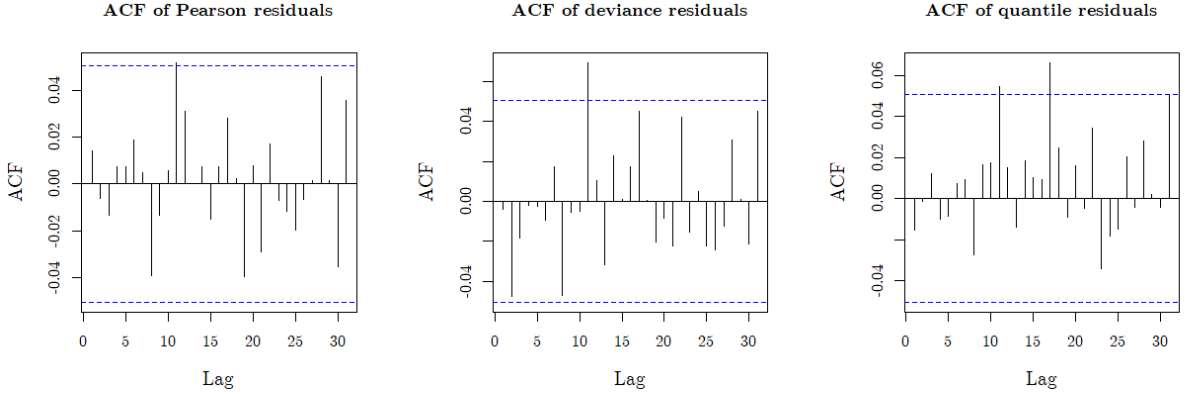

Figure B.4: Autocorrelation plot of Pearson, deviance and randomised quantile residuals in residuals for IC Scenario 2 (ZIP).

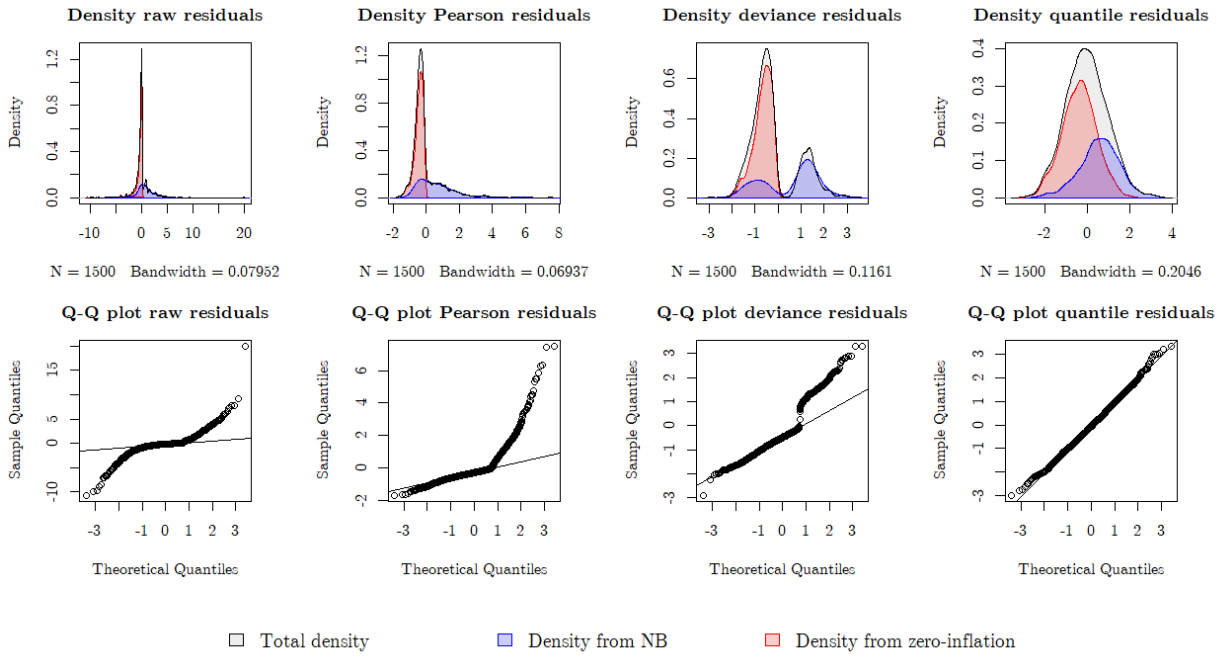

Figure B.5: Density and Q-Q plots of ZIP regression residuals for IC Scenario 3 (ZINB), indicating whether the residuals originates from the zero-inflation (red) or the Poisson distribution (blue).

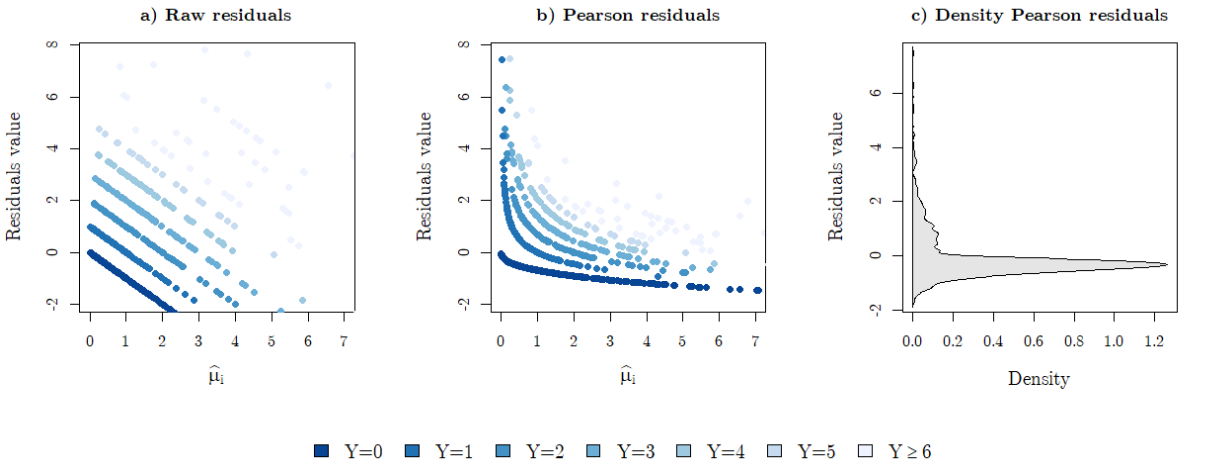

Figure B.6: Breakdown of Pearson residuals in IC Scenario 3 (ZINB), with: a) Raw residuals $r_j = y_j - \hat{\mu}_j$ plotted against prediction $\hat{\mu}_j$, b) Pearson residuals r_j^P plotted against prediction $\hat{\mu}_j$ and c) a density plot of Pearson residuals.

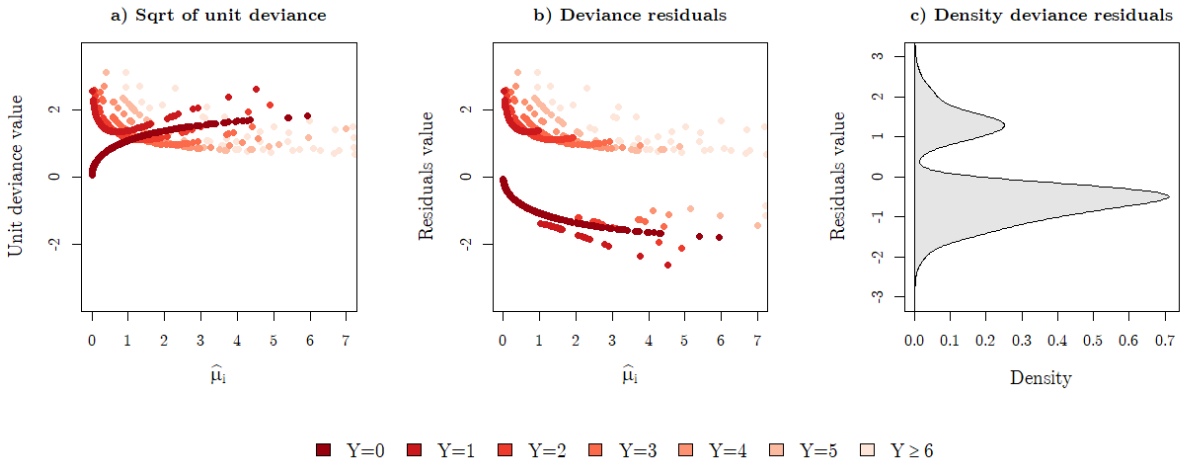

Figure B.7: Breakdown of deviance residuals in IC Scenario 3 (ZINB), with: a) the squared unit deviance $\sqrt{d_j}$ plotted against prediction $\hat{\mu}_j$, b) deviance residuals r_j^D plotted against prediction $\hat{\mu}_j$ and c) a density plot of deviance residuals.

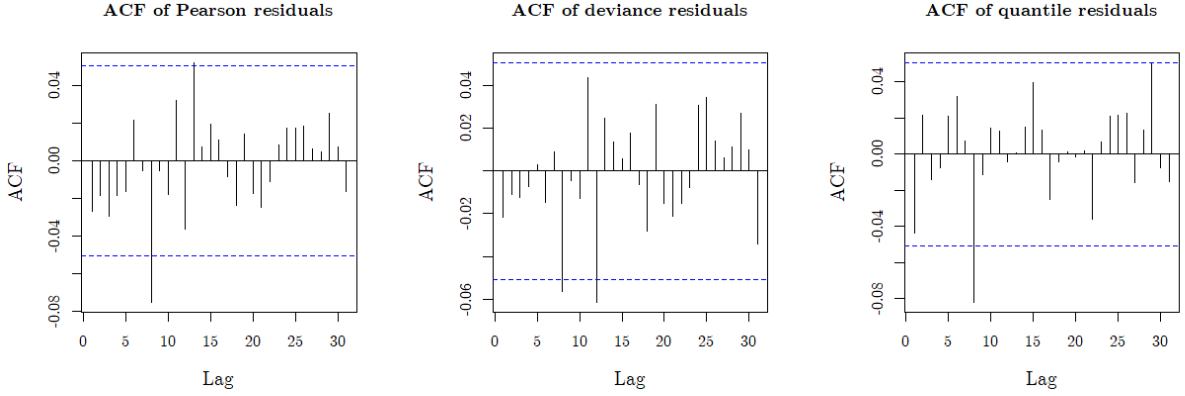

Figure B.8: Autocorrelation plot of Pearson, deviance and randomised quantile residuals in residuals for IC Scenario 3 (ZINB).

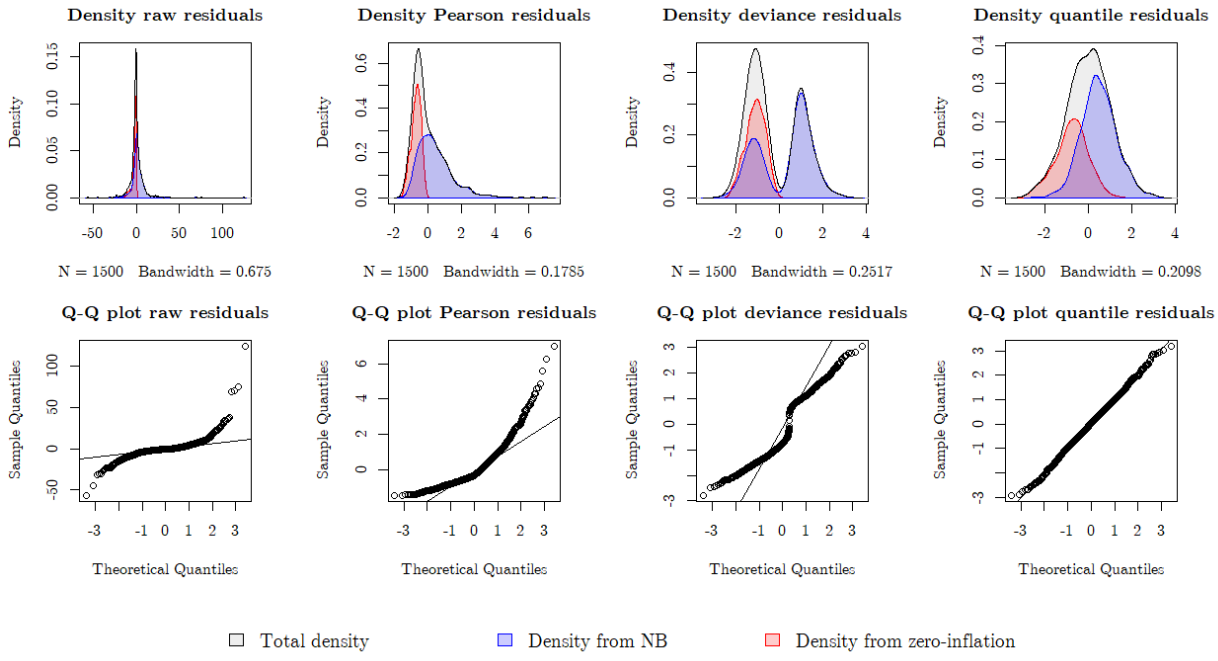

Figure B.9: Density and Q-Q plots of ZIP regression residuals for IC Scenario 4 (ZINB), indicating whether the residuals originates from the zero-inflation (red) or the Poisson distribution (blue).

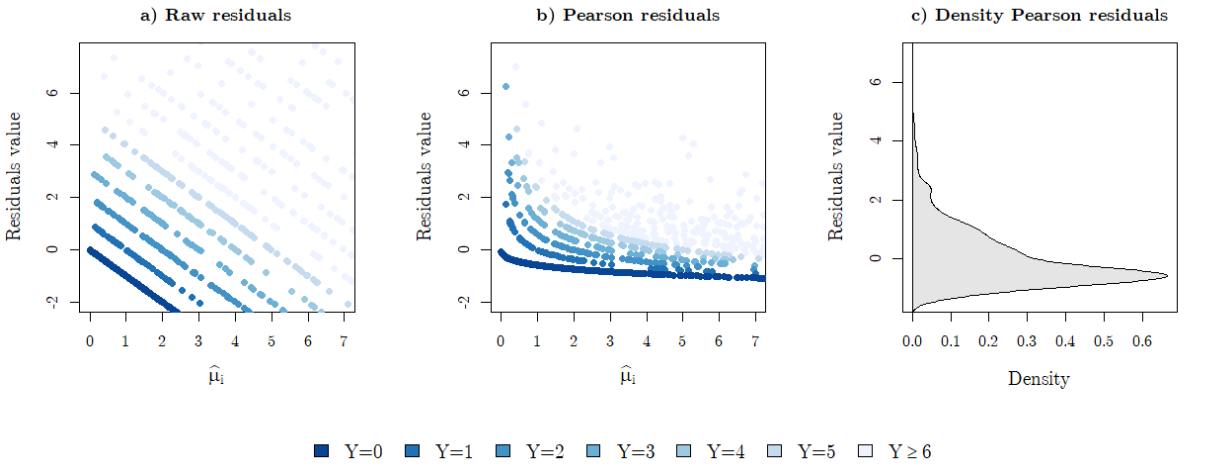

Figure B.10: Breakdown of Pearson residuals in IC Scenario 4 (ZINB), with: a) Raw residuals $r_j = y_j - \hat{\mu}_j$ plotted against prediction $\hat{\mu}_j$, b) Pearson residuals r_j^P plotted against prediction $\hat{\mu}_j$ and c) a density plot of Pearson residuals.

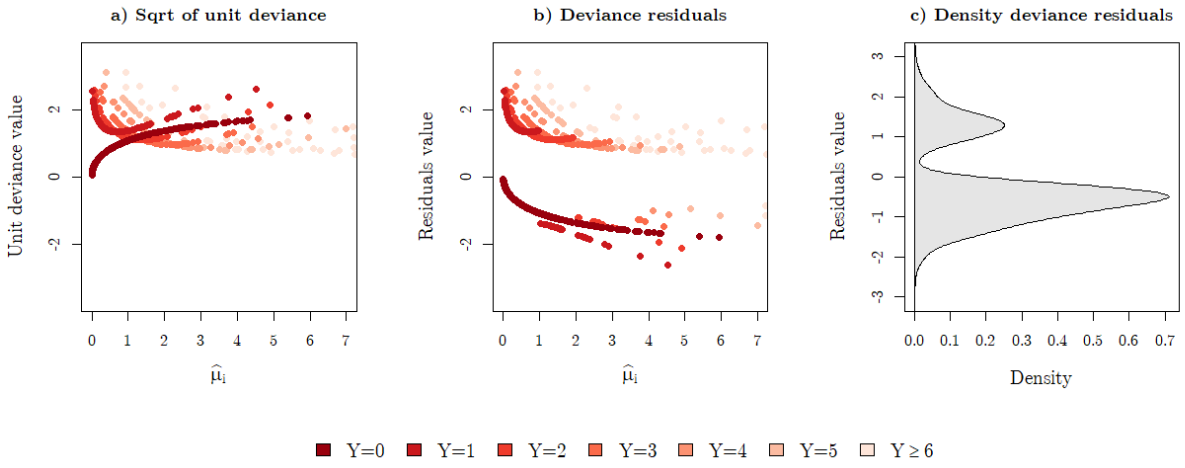

Figure B.11: Breakdown of deviance residuals in IC Scenario 4 (ZINB), with: a) the squared unit deviance $\sqrt{d_j}$ plotted against prediction $\hat{\mu}_j$, b) deviance residuals r_j^D plotted against prediction $\hat{\mu}_j$ and c) a density plot of deviance residuals.

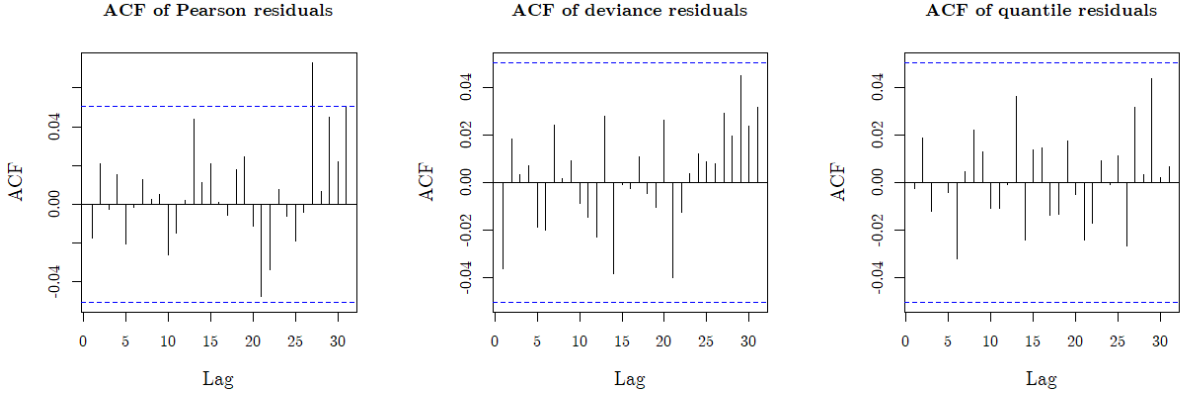

Figure B.12: Autocorrelation plot of Pearson, deviance and randomised quantile residuals in residuals for IC Scenario 4 (ZINB).

				No OC signal (% of total)		
IC scenario	Residuals	ARL_0	$SDRL_0$	$n = 2000$	$n = 3000$	$n = 4000$
ZIP 1	Pearson	197.92	202.47	0.00	0.00	0.00
ZIP 1	Deviance	193.09	197.36	0.00	0.00	0.00
ZIP 1	Quantile	195.66	194.08	0.01	0.00	0.00
ZIP 2	Pearson	203.14	205.00	0.01	0.00	0.00
ZIP 2	Deviance	198.89	201.49	0.01	0.00	0.00
ZIP 2	Quantile	202.03	204.90	0.00	0.00	0.00
ZINB 3	Pearson	197.98	202.93	0.02	0.00	0.00
ZINB 3	Deviance	192.91	200.62	0.01	0.00	0.00
ZINB 3	Quantile	196.84	199.18	0.00	0.00	0.00
ZINB 3	Pearson	202.39	207.40	0.02	0.00	0.00
ZINB 3	Deviance	199.44	205.67	0.01	0.00	0.00
ZINB 3	Quantile	198.49	202.19	0.01	0.00	0.00

Table B.2: Percentage of runs with no OC signal for the ZIP and ZINB regression-based Shewhart charts with symmetric control limits, for $n = 2000, 3000, 4000$.

				No OC signal (% of total)		
IC scenario	Residuals	ARL_0	$SDRL_0$	$n = 2000$	$n = 3000$	$n = 4000$
ZIP 1	Pearson	197.98	201.70	0.00	0.00	0.00
ZIP 1	Deviance	194.76	196.27	0.00	0.00	0.00
ZIP 1	Quantile	199.85	200.79	0.02	0.00	0.00
ZIP 2	Pearson	198.66	200.51	0.00	0.00	0.00
ZIP 2	Deviance	196.69	199.75	0.01	0.00	0.00
ZIP 2	Quantile	197.49	198.12	0.01	0.00	0.00
ZINB 3	Pearson	201.26	205.66	0.01	0.00	0.00
ZINB 3	Deviance	195.76	204.94	0.02	0.00	0.00
ZINB 3	Quantile	195.87	199.72	0.00	0.00	0.00
ZINB 4	Pearson	203.78	208.60	0.02	0.00	0.00
ZINB 4	Deviance	198.68	203.58	0.02	0.00	0.00
ZINB 4	Quantile	202.90	203.91	0.00	0.00	0.00

Table B.3: Percentage of runs with no OC signal for the ZIP and ZINB regression-based Shewhart charts with probability control limits, for $n = 2000, 3000, 4000$.

B.3 Baseline performance of regression-based Shewhart chart

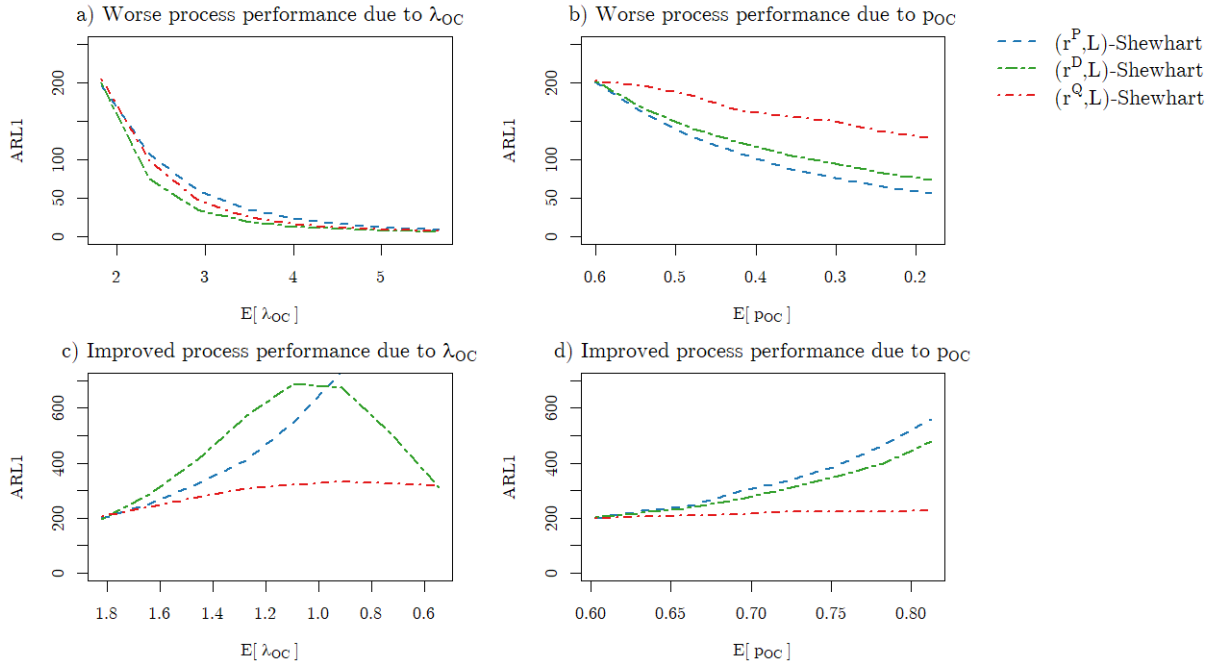

Figure B.13: Baseline ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart for IC ZINB Scenario 3 with $E[p^{IC}] = 0.61$ and $E[\lambda^{IC}] = 1.82$.

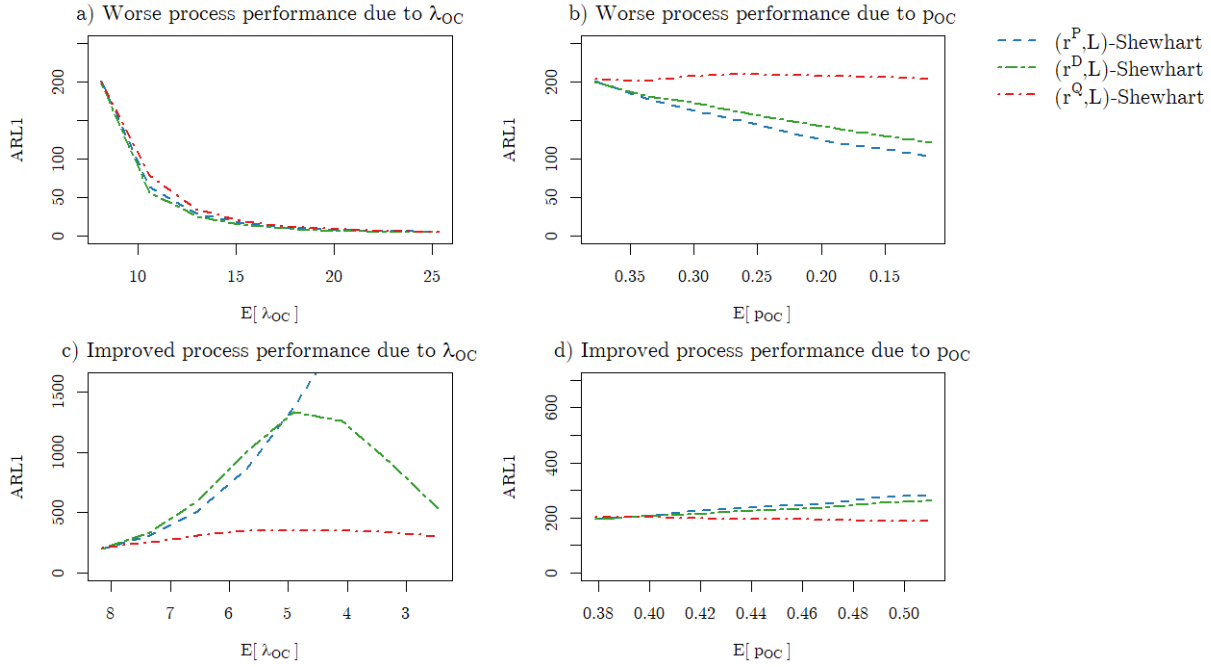

Figure B.14: Baseline ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart chart for IC ZINB Scenario 4 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.16$.

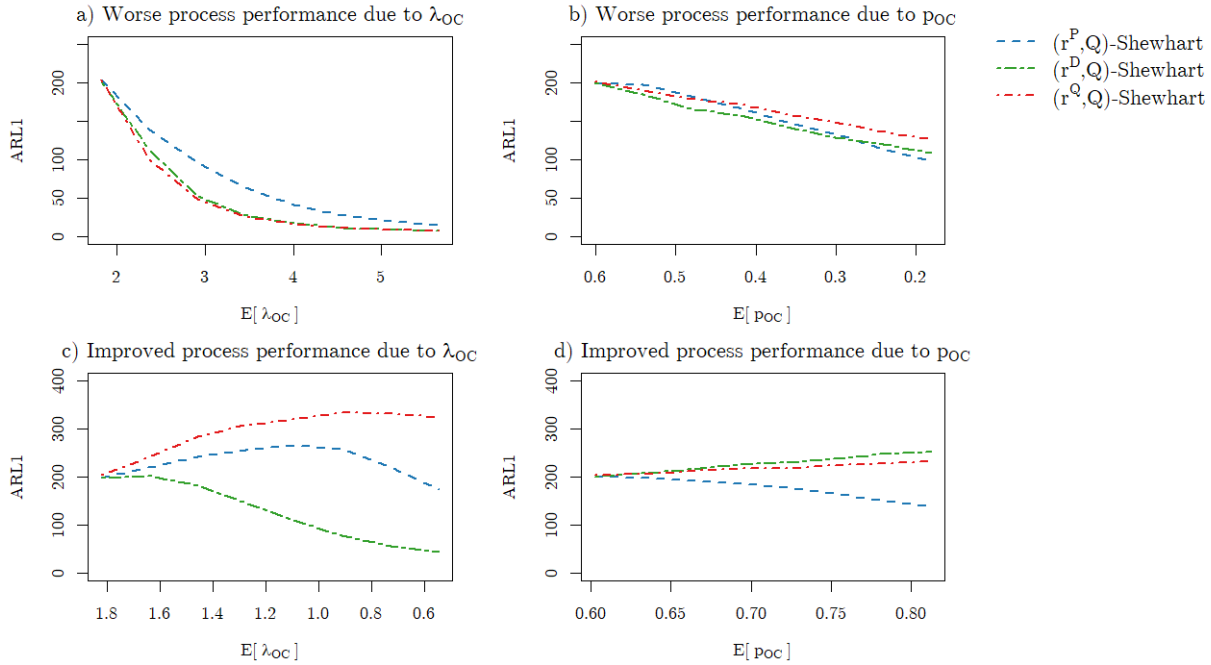

Figure B.15: Baseline ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart for IC ZINB Scenario 3 with $E[p^{IC}] = 0.61$ and $E[\lambda^{IC}] = 1.82$.

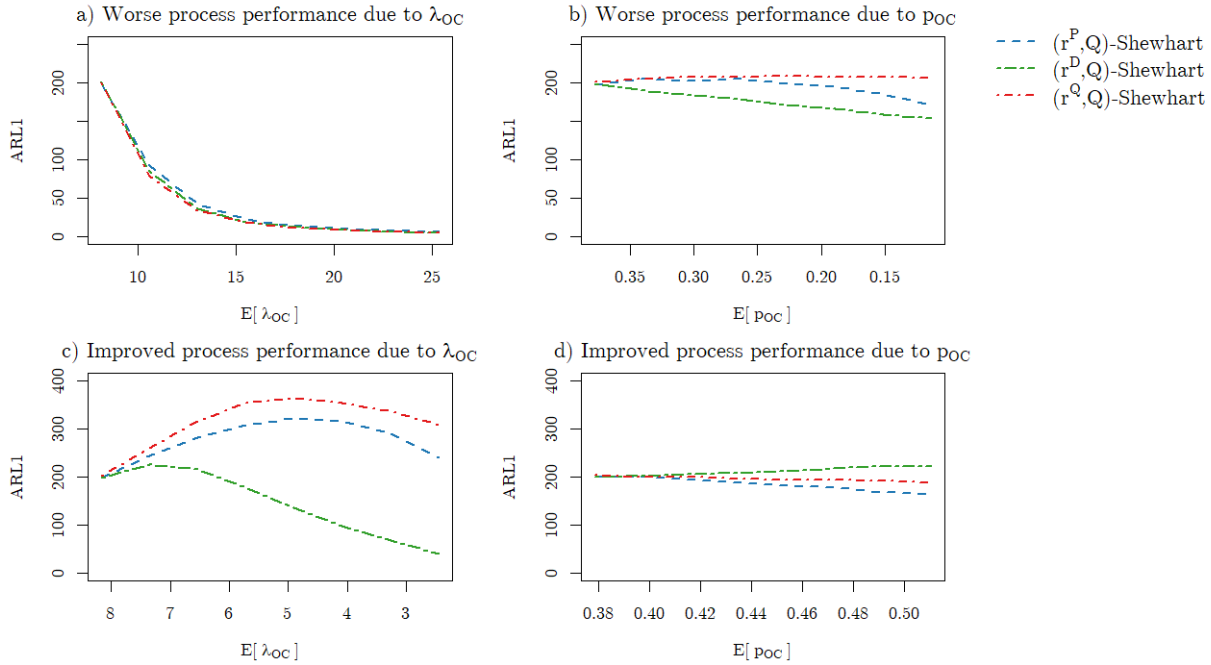

Figure B.16: Baseline ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart chart for IC ZINB Scenario 4 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.16$.

B.4 Performance results while considering Phase I effects

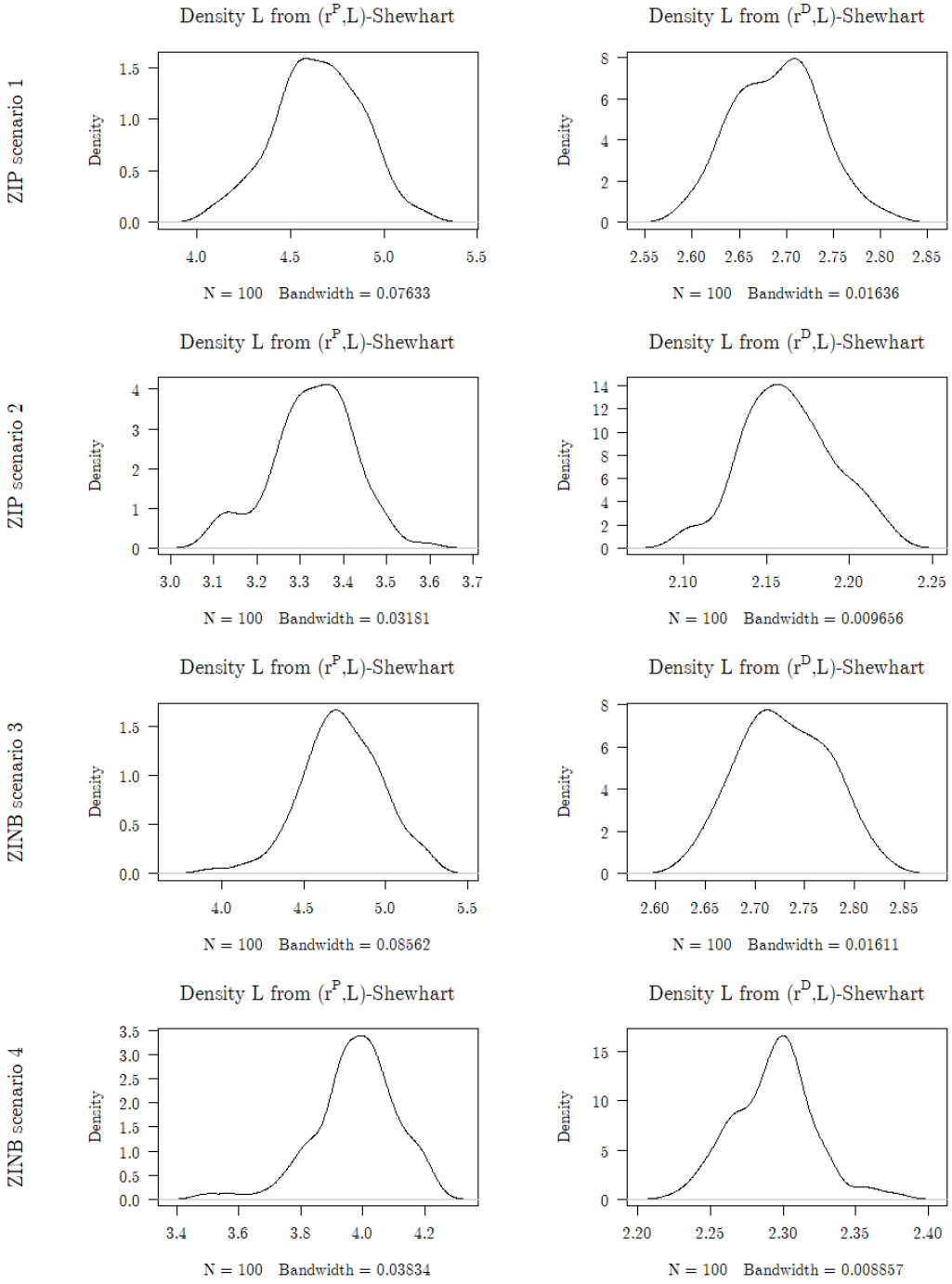

Figure B.17: The density of the charting constant L, solved 100 times for performance evaluation while taking into account the effects of Phase I, and for the (r^P, L)- and (r^D, L)-Shewhart chart in each IC scenario.

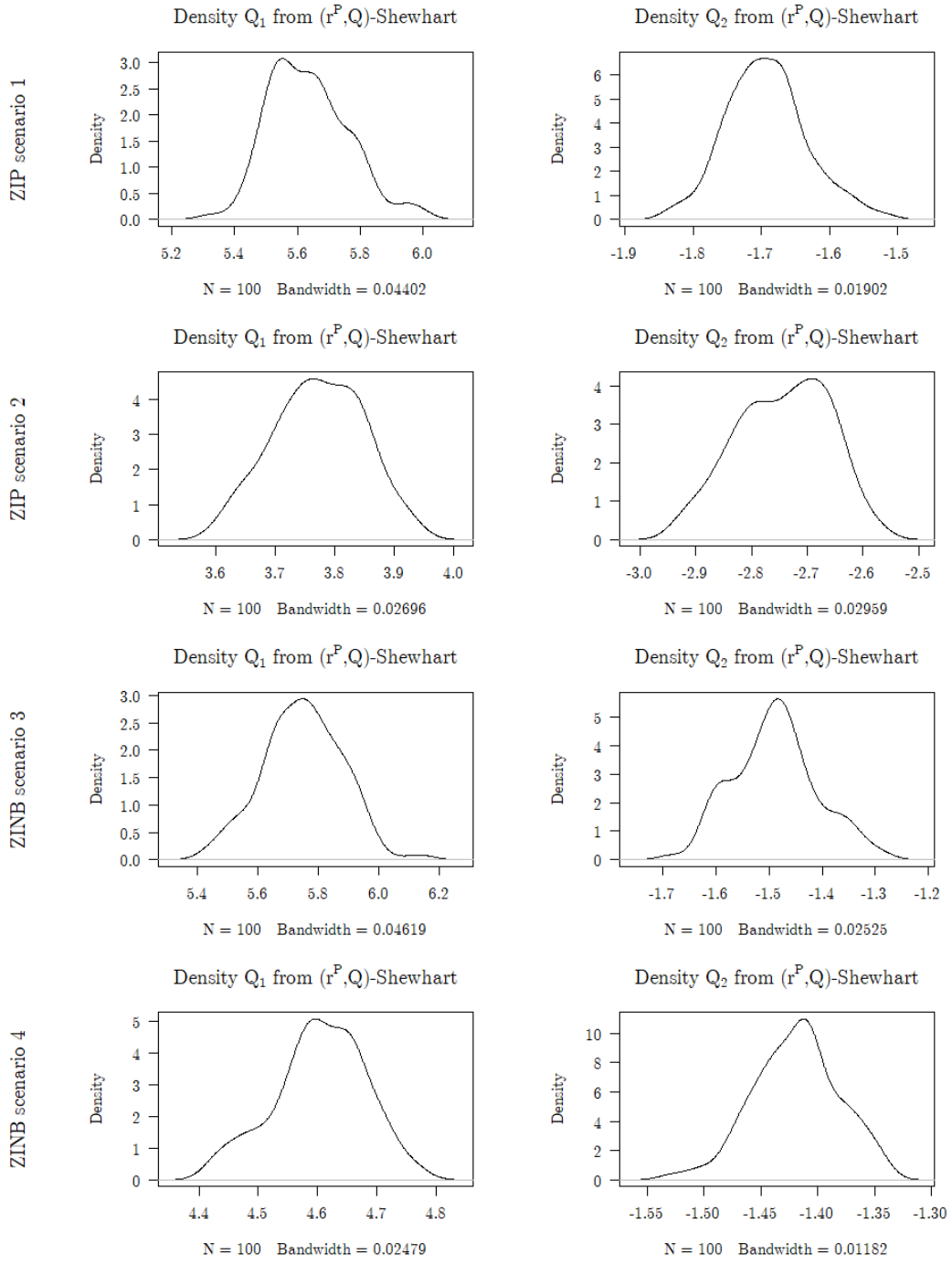

Figure B.18: The density of the charting constants Q_1 and Q_2, solved 100 times for performance evaluation while taking into account the effects of Phase I, and for the (r^P, Q)-Shewhart chart in each IC scenario.

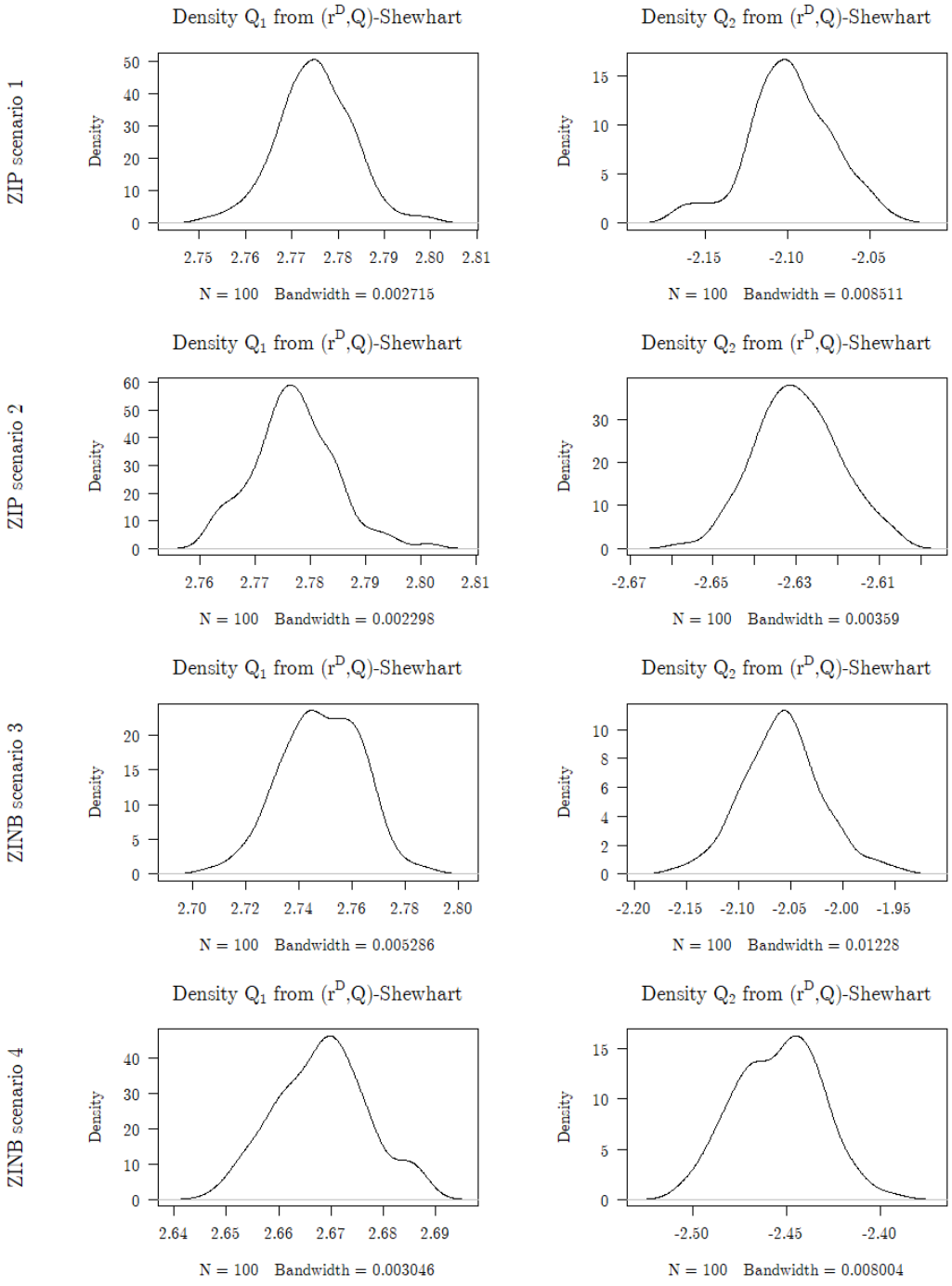

Figure B.19: The density of the charting constants Q_1 and Q_2, solved 100 times for performance evaluation while taking into account the effects of Phase I, and for the (r^D, Q)-Shewhart chart in each IC scenario.

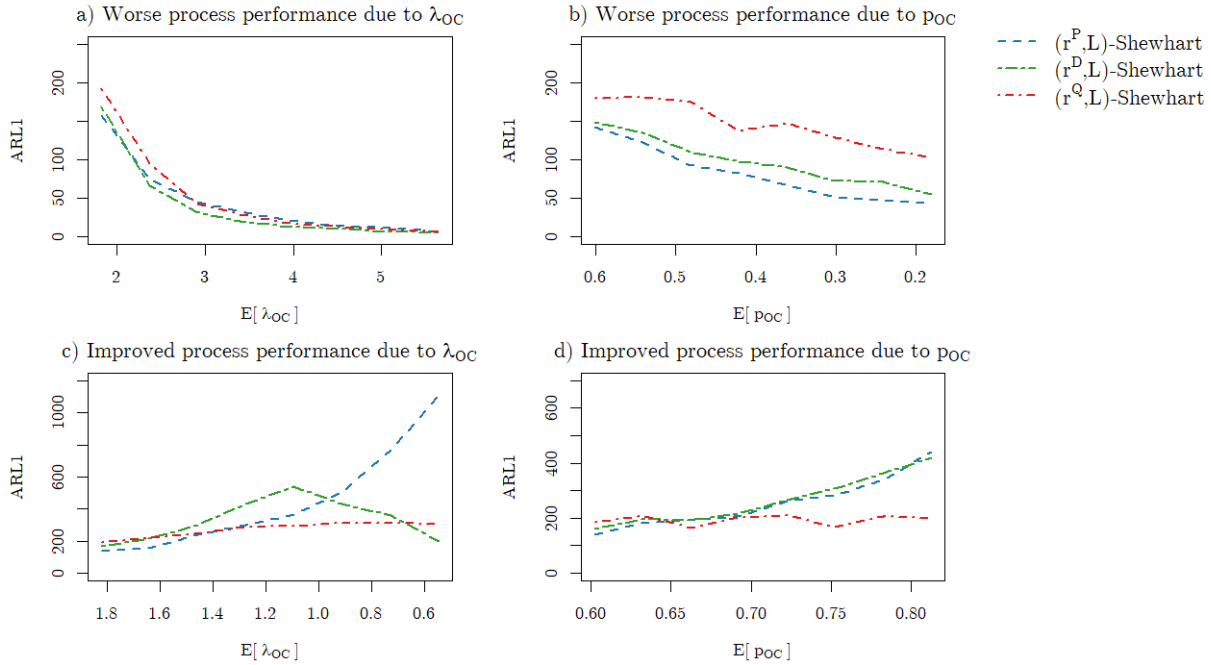

Figure B.20: ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart, while considering Phase I estimates, for IC ZINB Scenario 3 with $E[p^{IC}] = 0.61$ and $E[\lambda^{IC}] = 1.82$.

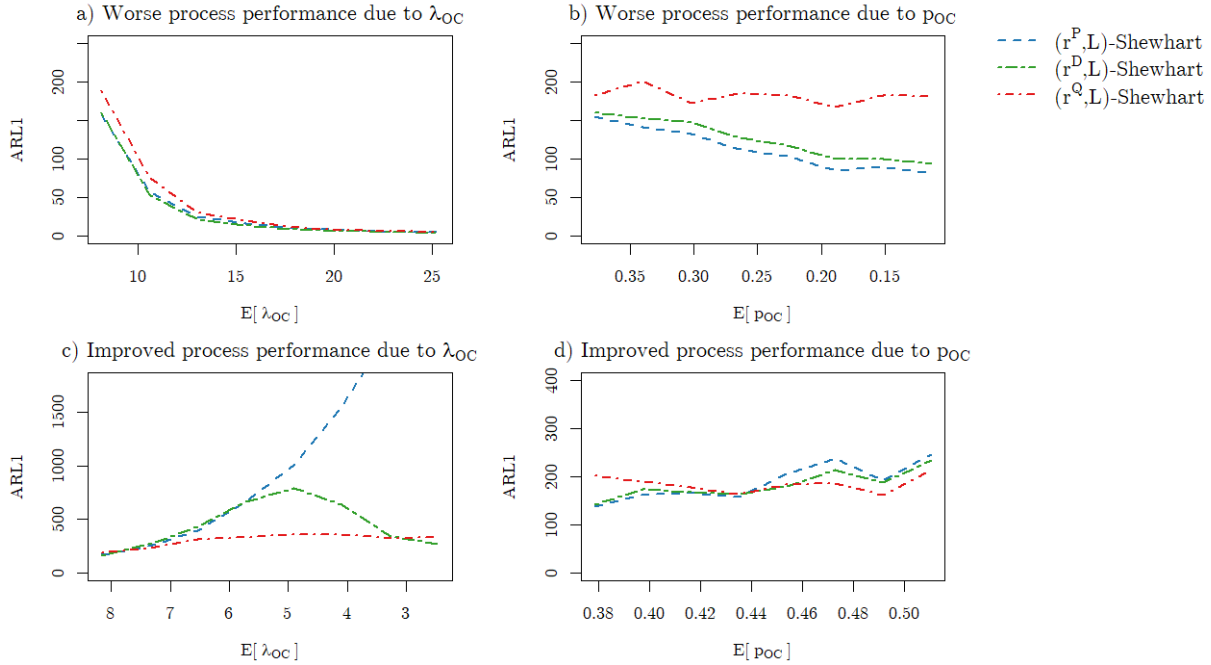

Figure B.21: ARL_1 performance of the (r^P, L)-, (r^D, L)- and (r^Q, L)-Shewhart, while considering Phase I estimates, for IC ZINB Scenario 4 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.16$.

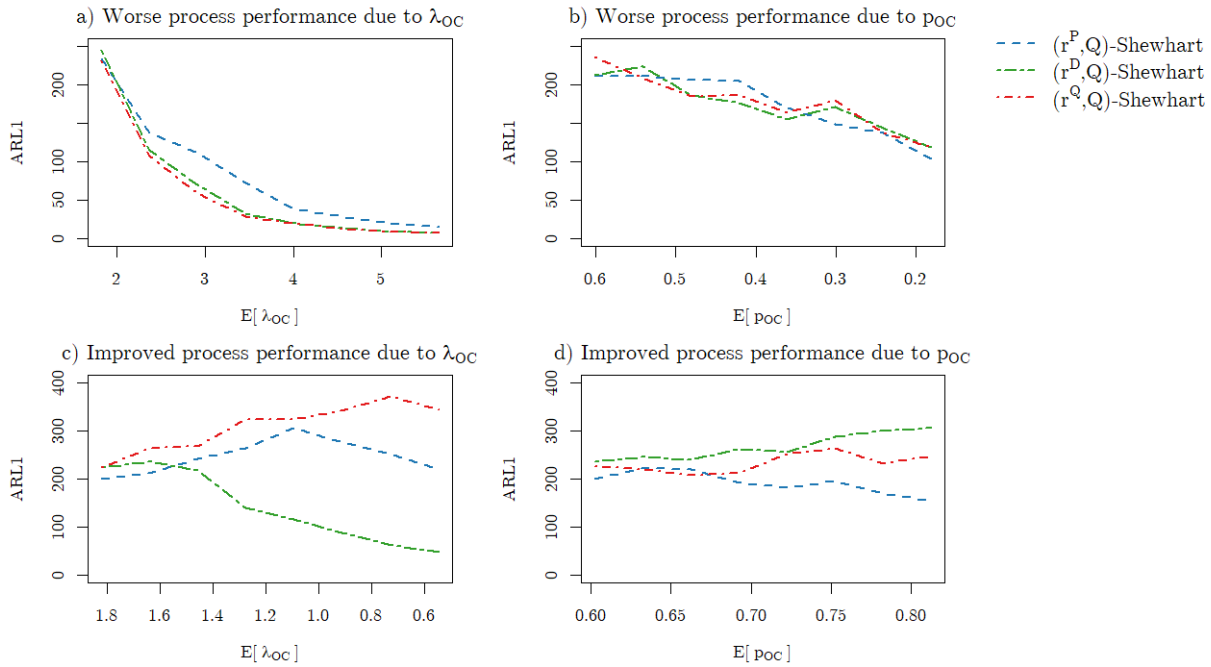

Figure B.22: ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart, while considering Phase I estimates, for IC ZINB Scenario 3 with $E[p^{IC}] = 0.61$ and $E[\lambda^{IC}] = 1.82$.

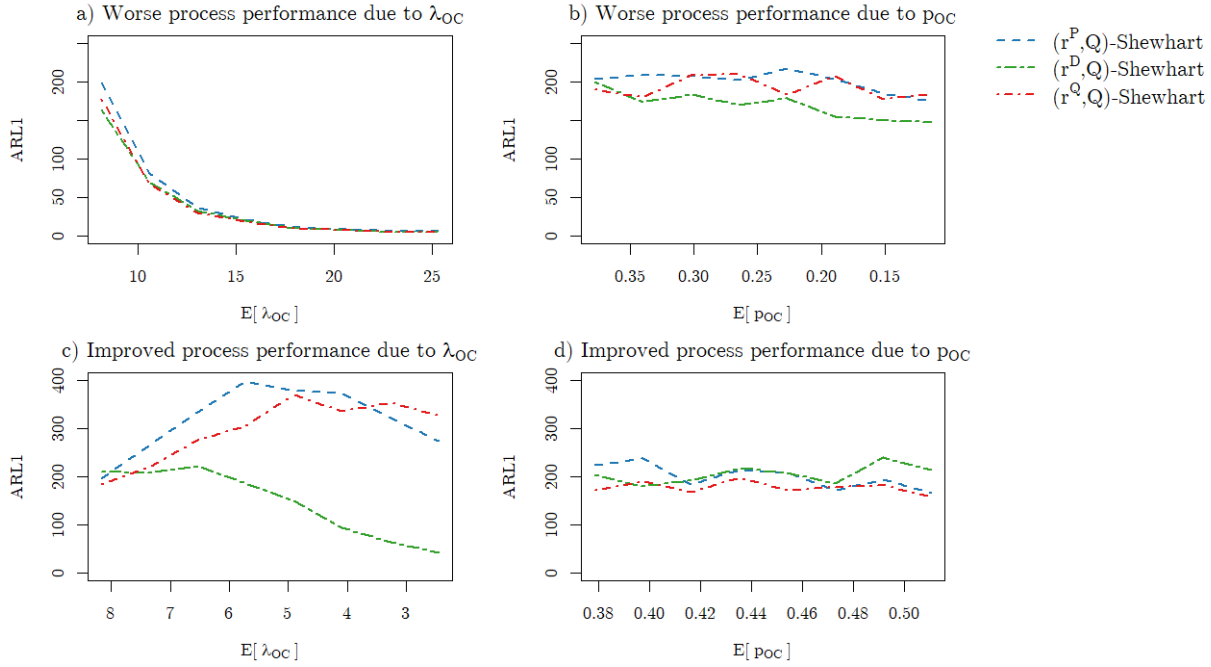

Figure B.23: ARL_1 performance of the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-Shewhart, while considering Phase I estimates, for IC ZINB Scenario 4 with $E[p^{OC}] = 0.38$ and $E[\lambda^{IC}] = 8.16$.

B.5 ARL and SDRL results of the regression-based Shewhart charts

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.10	1.82	197.87	197.78	196.23	195.00	205.37	206.83
-0.01	1.64	252.19	255.48	279.91	278.15	246.23	246.97
-0.12	1.46	307.01	308.33	370.11	372.49	274.67	272.31
-0.26	1.28	401.10	402.18	403.93	398.42	306.12	305.46
-0.41	1.09	524.25	524.96	322.63	321.75	318.18	315.73
-0.59	0.91	693.62	688.94	204.47	204.42	332.53	333.42
-0.82	0.73	983.50	913.51	129.59	127.76	332.20	336.41
-1.10	0.55	1398.71	1182.04	85.99	86.28	318.78	318.16
0.10	1.82	202.33	202.66	199.73	198.15	201.62	203.68
0.36	2.37	110.18	108.96	66.51	65.56	91.50	92.01
0.57	2.92	66.61	66.17	27.64	27.32	38.32	37.16
0.74	3.46	39.84	39.62	14.61	14.37	19.52	18.96
0.89	4.01	25.29	25.06	9.37	8.99	11.80	11.46
1.02	4.56	15.92	15.20	7.07	6.48	8.53	7.91
1.13	5.10	10.59	10.02	5.58	5.01	6.47	5.92
1.23	5.65	7.82	7.29	4.74	4.16	5.42	4.81

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.50	0.60	198.30	198.83	200.07	199.53	204.75	203.48
0.20	0.54	162.14	162.85	169.20	166.83	195.00	199.46
-0.09	0.48	132.85	133.10	143.42	142.16	179.85	178.52
-0.38	0.42	106.39	105.94	122.64	122.84	172.09	171.10
-0.69	0.36	87.75	86.71	105.15	104.40	156.42	156.40
-1.01	0.30	76.52	77.11	95.08	93.96	149.45	148.74
-1.37	0.24	63.89	63.49	84.75	84.94	139.89	139.56
-1.79	0.18	55.10	54.37	75.36	74.33	128.08	126.28
0.50	0.60	195.08	194.66	195.14	195.47	200.82	203.05
0.65	0.63	229.40	230.93	216.77	214.34	208.10	208.72
0.81	0.66	254.78	255.34	240.55	246.33	206.66	204.92
0.98	0.69	293.89	294.37	265.96	266.70	218.13	215.24
1.15	0.72	341.04	337.38	303.48	307.97	217.01	218.17
1.33	0.75	405.91	409.66	348.98	350.64	224.00	224.42
1.53	0.78	466.72	474.64	395.96	402.82	227.54	229.45
1.74	0.81	578.17	577.00	462.56	457.85	228.62	227.56

Table B.4: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
1.60	8.17	200.03	200.96	198.17	194.30	199.92	200.23
1.49	7.35	266.14	263.79	219.36	223.59	277.85	278.05
1.38	6.53	356.75	360.77	112.69	113.69	336.67	345.71
1.24	5.72	449.94	448.38	45.97	45.56	358.54	361.55
1.09	4.90	564.09	564.93	21.29	20.87	367.81	363.85
0.91	4.08	597.51	591.00	11.85	11.40	366.74	363.06
0.68	3.27	498.76	491.07	7.72	7.19	344.79	345.40
0.40	2.45	309.26	308.30	5.54	5.02	323.80	328.90
1.60	8.17	202.57	199.98	200.37	198.91	203.92	202.43
1.86	10.62	81.92	80.55	24.82	24.75	30.64	30.61
2.07	13.07	27.00	26.48	7.11	6.67	8.36	7.86
2.24	15.52	9.29	8.72	3.85	3.34	4.28	3.74
2.39	17.97	4.58	4.08	2.80	2.23	3.02	2.47
2.52	20.42	3.00	2.45	2.30	1.73	2.43	1.87
2.63	22.87	2.33	1.76	2.05	1.48	2.13	1.58
2.73	25.31	2.03	1.46	1.91	1.32	1.96	1.37

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
-0.60	0.38	202.79	200.84	201.71	203.40	200.69	200.35
-0.80	0.34	182.40	181.69	185.67	186.31	203.02	201.16
-1.00	0.30	164.86	163.01	172.48	168.96	209.50	209.83
-1.22	0.26	151.62	150.93	165.84	164.86	208.89	210.51
-1.46	0.23	137.62	136.12	158.13	156.94	213.17	211.81
-1.73	0.19	125.05	124.54	149.18	146.33	215.45	212.79
-2.04	0.15	109.99	110.51	140.18	140.02	216.51	215.79
-2.41	0.11	98.45	97.62	131.84	132.39	209.50	208.21
-0.60	0.38	202.49	201.17	199.17	197.97	198.62	198.24
-0.50	0.40	207.37	207.19	208.02	208.66	198.61	197.40
-0.41	0.42	215.62	214.08	215.79	217.90	200.74	200.03
-0.32	0.43	221.32	222.04	220.63	218.39	194.43	191.14
-0.22	0.45	232.36	231.16	232.81	233.25	192.44	190.37
-0.13	0.47	236.53	237.69	237.24	238.05	191.29	189.30
-0.04	0.49	244.70	244.04	247.78	245.00	185.30	181.07
0.05	0.51	249.52	246.39	259.14	261.01	183.66	180.73

Table B.5: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, for IC ZIP Scenario 2 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.10	1.82	198.08	194.59	194.92	190.84	204.54	202.71
-0.01	1.64	250.81	248.42	286.91	282.86	240.67	243.71
-0.12	1.46	323.50	320.38	414.26	416.24	276.43	274.29
-0.26	1.28	409.60	407.07	570.26	563.56	308.17	308.48
-0.41	1.09	548.21	547.25	687.51	675.64	322.36	325.62
-0.59	0.91	734.39	712.59	674.36	662.76	332.25	335.06
-0.82	0.73	1026.11	958.13	505.37	505.24	325.60	321.70
-1.10	0.55	1484.99	1229.41	309.66	304.92	318.92	325.69
0.10	1.82	196.23	197.23	199.30	200.44	204.68	205.78
0.36	2.37	107.04	105.52	75.24	74.34	98.14	94.90
0.57	2.92	59.90	58.28	34.77	34.01	47.67	47.52
0.74	3.46	36.67	36.42	19.69	19.09	26.72	26.23
0.89	4.01	23.14	22.44	12.70	12.29	16.41	15.98
1.02	4.56	15.73	15.14	9.32	8.80	11.49	10.85
1.13	5.10	11.54	10.85	7.29	6.71	8.80	8.24
1.23	5.65	9.09	8.59	6.16	5.61	7.27	6.70

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.50	0.60	201.59	200.30	202.51	204.07	202.11	201.13
0.20	0.54	160.91	161.67	167.13	166.39	195.98	197.27
-0.09	0.48	130.17	130.11	141.66	140.97	184.01	184.67
-0.38	0.42	108.72	110.61	121.83	122.30	163.40	163.04
-0.69	0.36	88.68	87.61	106.22	105.43	155.87	156.60
-1.01	0.30	76.05	74.02	95.03	93.40	149.72	148.07
-1.37	0.24	64.30	64.52	82.54	83.64	136.35	137.52
-1.79	0.18	55.59	54.49	73.56	71.86	127.41	128.26
0.50	0.60	200.41	197.27	201.39	200.32	200.74	201.94
0.65	0.63	226.80	229.96	219.52	217.57	205.83	204.84
0.81	0.66	246.09	247.81	237.21	240.69	211.36	214.30
0.98	0.69	296.92	298.53	270.56	267.88	213.33	216.59
1.15	0.72	336.19	335.82	307.61	308.50	223.77	225.44
1.33	0.75	389.75	384.79	348.62	353.51	224.22	224.97
1.53	0.78	465.17	473.92	399.38	401.88	223.00	225.39
1.74	0.81	560.72	555.57	478.96	475.48	228.35	227.59

Table B.6: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, for IC ZINB Scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
1.60	8.17	199.44	199.76	198.81	196.62	203.74	205.38
1.49	7.35	313.53	315.41	337.89	335.78	258.49	258.90
1.38	6.53	511.67	504.02	600.90	599.17	312.56	311.24
1.24	5.72	854.24	820.70	1000.32	930.82	348.54	342.19
1.09	4.90	1372.18	1159.06	1331.88	1146.28	350.20	348.03
0.91	4.08	2005.68	1390.28	1264.90	1109.14	349.25	353.41
0.68	3.27	2643.65	1427.74	914.42	872.50	334.84	329.63
0.40	2.45	3234.63	1233.03	529.94	530.56	304.28	302.48
1.60	8.17	200.66	199.95	197.72	196.51	199.39	198.70
1.86	10.62	63.71	62.74	56.03	55.92	78.34	78.49
2.07	13.07	28.25	27.91	24.44	24.22	33.49	33.07
2.24	15.52	15.87	15.35	13.90	13.40	18.50	18.01
2.39	17.97	10.28	9.83	9.15	8.59	11.74	11.32
2.52	20.42	7.44	6.94	6.73	6.21	8.32	7.83
2.63	22.87	5.83	5.32	5.39	4.92	6.44	5.89
2.73	25.31	4.85	4.34	4.48	3.94	5.27	4.77

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, L)-Shewhart		(r^D, L)-Shewhart		(r^Q, L)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
-0.60	0.38	201.54	201.58	200.20	199.02	203.10	203.25
-0.80	0.34	180.20	178.79	182.77	179.47	201.04	198.74
-1.00	0.30	163.84	161.23	173.04	171.81	207.52	204.36
-1.22	0.26	149.73	149.78	161.42	160.00	210.00	210.48
-1.46	0.23	134.71	135.43	149.78	150.30	208.25	206.27
-1.73	0.19	121.06	120.83	139.27	139.60	207.68	204.19
-2.04	0.15	113.24	114.67	130.18	130.85	205.58	203.31
-2.41	0.11	102.48	101.36	121.57	120.17	203.48	204.25
-0.60	0.38	196.70	197.16	195.58	197.64	201.33	202.97
-0.50	0.40	206.54	204.23	205.25	204.95	202.85	199.52
-0.41	0.42	222.15	225.90	213.25	214.37	198.62	198.85
-0.32	0.43	232.48	234.81	224.96	224.99	196.18	193.69
-0.22	0.45	244.96	243.78	232.08	230.83	195.19	196.57
-0.13	0.47	256.34	254.83	241.18	237.66	191.80	191.22
-0.04	0.49	275.60	270.74	254.53	253.11	188.66	188.35
0.05	0.51	283.45	283.27	261.63	261.69	189.62	189.08

Table B.7: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, for IC ZINB Scenario 4 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.10	1.82	198.43	200.01	196.40	199.20	199.63	199.11
-0.01	1.64	218.76	219.25	181.91	178.96	240.01	237.98
-0.12	1.46	240.09	240.73	141.32	140.62	278.93	278.09
-0.26	1.28	258.38	258.38	95.84	96.57	309.18	315.64
-0.41	1.09	264.94	264.52	67.38	66.50	318.30	318.42
-0.59	0.91	240.44	240.33	46.39	46.94	335.12	335.15
-0.82	0.73	200.04	199.41	34.38	33.35	328.21	321.96
-1.10	0.55	150.26	149.43	27.94	27.36	323.88	326.14
0.10	1.82	198.76	195.64	199.61	201.54	203.78	207.01
0.36	2.37	145.54	143.25	104.21	104.27	91.34	90.72
0.57	2.92	101.94	100.89	40.50	40.22	38.73	38.36
0.74	3.46	69.85	69.97	20.00	19.22	19.63	18.81
0.89	4.01	47.14	46.64	11.95	11.40	11.84	11.26
1.02	4.56	31.94	31.56	8.48	7.98	8.48	8.04
1.13	5.10	21.16	20.79	6.58	6.02	6.58	6.04
1.23	5.65	14.13	13.70	5.42	4.92	5.44	4.94

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.50	0.60	200.29	200.67	199.52	196.55	202.99	201.23
0.20	0.54	194.40	194.43	185.05	184.62	191.02	190.48
-0.09	0.48	183.12	181.10	163.56	165.37	178.56	177.90
-0.38	0.42	164.53	163.67	150.10	151.42	166.85	164.98
-0.69	0.36	147.08	145.98	139.82	139.21	159.86	162.68
-1.01	0.30	130.26	128.81	126.28	124.73	149.21	150.86
-1.37	0.24	113.16	111.34	118.05	115.00	137.18	135.63
-1.79	0.18	95.51	94.44	106.41	108.21	125.25	126.09
0.50	0.60	201.13	197.42	201.37	199.95	202.31	200.68
0.65	0.63	197.43	196.11	207.07	211.48	205.10	204.40
0.81	0.66	194.07	193.34	221.90	220.23	210.70	209.72
0.98	0.69	187.40	185.12	232.86	227.82	215.43	214.67
1.15	0.72	176.39	174.90	244.21	243.13	218.82	220.63
1.33	0.75	164.46	164.66	257.18	259.97	221.27	221.12
1.53	0.78	150.00	151.80	262.80	262.06	226.01	229.16
1.74	0.81	135.96	134.56	272.29	271.62	226.32	224.83

Table B.8: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, for IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
1.60	8.17	199.21	193.47	202.88	202.44	202.93	204.95
1.49	7.35	233.50	231.62	138.65	137.29	280.00	280.12
1.38	6.53	266.93	262.93	63.35	63.35	330.54	327.37
1.24	5.72	294.39	294.07	28.03	27.60	361.76	360.42
1.09	4.90	292.09	284.47	14.64	14.18	356.88	355.33
0.91	4.08	241.08	243.02	8.80	8.32	363.83	364.41
0.68	3.27	164.91	163.58	5.98	5.43	348.01	350.56
0.40	2.45	102.21	101.64	4.56	4.05	325.00	326.68
1.60	8.17	202.03	202.86	201.16	199.01	202.74	200.29
1.86	10.62	111.36	111.40	32.81	32.02	31.53	30.96
2.07	13.07	47.04	46.33	8.34	7.79	8.28	7.71
2.24	15.52	15.86	15.48	4.32	3.76	4.30	3.76
2.39	17.97	6.76	6.23	3.03	2.43	3.02	2.43
2.52	20.42	3.85	3.31	2.42	1.85	2.43	1.86
2.63	22.87	2.81	2.32	2.12	1.56	2.12	1.55
2.73	25.31	2.26	1.67	1.95	1.35	1.95	1.36

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
-0.60	0.38	197.87	199.74	202.20	203.68	200.45	199.59
-0.80	0.34	200.41	200.80	186.61	183.54	204.53	204.25
-1.00	0.30	200.46	200.75	182.87	182.60	208.83	206.77
-1.22	0.26	195.33	194.31	170.34	167.03	208.87	208.81
-1.46	0.23	185.13	180.24	163.34	161.18	211.57	210.41
-1.73	0.19	177.89	177.96	157.32	153.39	217.15	217.32
-2.04	0.15	161.90	159.75	147.49	145.24	212.64	213.16
-2.41	0.11	150.91	150.04	146.10	147.41	213.08	211.39
-0.60	0.38	197.40	199.93	200.76	200.22	200.30	196.65
-0.50	0.40	195.88	195.95	208.94	209.10	202.42	203.44
-0.41	0.42	193.26	193.42	210.93	211.87	197.07	197.45
-0.32	0.43	190.90	188.41	221.43	223.40	195.59	193.77
-0.22	0.45	188.55	188.30	220.52	217.46	190.54	189.17
-0.13	0.47	183.02	182.03	230.98	229.10	189.09	189.03
-0.04	0.49	172.50	174.78	240.64	238.57	188.34	186.93
0.05	0.51	168.13	166.59	244.22	244.40	183.94	183.39

Table B.9: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, for IC ZIP Scenario 2 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.10	1.82	198.67	198.80	198.49	200.97	203.64	211.16
-0.01	1.64	220.49	220.57	201.44	202.10	242.48	239.64
-0.12	1.46	240.83	242.38	181.06	180.66	283.87	282.76
-0.26	1.28	255.71	254.95	146.08	145.12	307.61	314.91
-0.41	1.09	266.63	266.42	109.84	106.38	319.22	314.96
-0.59	0.91	257.80	257.95	78.27	76.11	332.81	332.88
-0.82	0.73	219.75	220.88	56.81	55.79	331.04	328.13
-1.10	0.55	173.71	174.08	42.90	42.51	323.93	321.11
0.10	1.82	203.27	203.64	200.72	203.92	204.06	201.11
0.36	2.37	139.02	139.14	112.26	110.55	99.69	99.75
0.57	2.92	96.31	96.08	52.72	52.35	48.85	48.57
0.74	3.46	63.03	62.71	27.89	27.17	26.46	25.87
0.89	4.01	41.29	41.02	17.13	16.51	16.53	15.86
1.02	4.56	27.62	27.15	11.58	11.20	11.38	11.00
1.13	5.10	19.36	18.87	8.86	8.39	8.74	8.31
1.23	5.65	14.40	13.85	7.18	6.67	7.14	6.67

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
0.50	0.60	198.44	198.44	200.16	199.95	200.55	199.09
0.20	0.54	197.48	197.24	184.94	185.21	189.99	192.33
-0.09	0.48	182.76	182.16	166.17	166.50	178.96	177.06
-0.38	0.42	167.83	167.44	157.20	156.58	172.09	169.99
-0.69	0.36	149.11	150.74	142.26	142.89	158.49	159.37
-1.01	0.30	133.91	134.40	128.83	128.32	148.99	149.53
-1.37	0.24	113.07	110.84	120.44	119.04	135.94	134.19
-1.79	0.18	98.01	96.58	109.18	108.75	126.61	125.03
0.50	0.60	199.05	196.78	199.70	199.35	204.05	202.10
0.65	0.63	198.62	199.28	206.82	206.18	205.71	204.11
0.81	0.66	192.14	192.21	216.23	218.52	212.11	213.18
0.98	0.69	186.80	183.86	225.13	227.92	216.98	217.76
1.15	0.72	178.00	177.78	229.76	223.96	217.42	213.52
1.33	0.75	165.80	164.83	238.63	241.79	224.59	228.49
1.53	0.78	150.39	149.44	248.06	243.49	227.51	221.03
1.74	0.81	137.11	137.19	251.10	252.13	232.41	236.12

Table B.10: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, for IC ZINB Scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
1.60	8.17	198.42	192.76	198.04	197.20	201.17	201.71
1.49	7.35	242.99	237.80	225.84	225.40	258.99	260.56
1.38	6.53	282.50	277.04	215.02	209.61	314.92	314.74
1.24	5.72	308.04	302.98	178.01	179.79	352.82	347.40
1.09	4.90	321.14	322.62	135.64	133.85	362.80	360.99
0.91	4.08	316.23	317.01	98.43	100.37	354.58	352.45
0.68	3.27	290.59	292.17	67.00	66.20	335.30	333.09
0.40	2.45	239.61	235.73	40.61	40.48	307.94	314.17
1.60	8.17	198.86	197.75	200.52	199.44	199.88	199.09
1.86	10.62	91.76	91.19	85.37	82.91	78.49	77.16
2.07	13.07	42.82	41.76	35.96	35.29	34.05	33.33
2.24	15.52	22.82	22.06	19.14	18.48	18.56	17.72
2.39	17.97	14.17	13.54	11.98	11.31	11.67	11.02
2.52	20.42	9.89	9.41	8.53	8.10	8.40	7.95
2.63	22.87	7.52	7.10	6.56	5.98	6.45	5.87
2.73	25.31	6.07	5.54	5.32	4.72	5.27	4.68

		OC scenario due to decreased/increased $E[p^{OC}]$					
		(r^P, Q)-Shewhart		(r^D, Q)-Shewhart		(r^Q, Q)-Shewhart	
γ_0^{OC}	$E[p^{OC}]$	ARL_1	$SDRL$	ARL_1	$SDRL$	ARL_1	$SDRL$
-0.60	0.38	197.42	192.85	198.90	197.84	201.69	201.16
-0.80	0.34	204.57	202.45	190.28	187.11	204.38	207.65
-1.00	0.30	202.07	199.18	183.05	182.32	207.20	204.29
-1.22	0.26	205.45	204.25	178.11	178.04	206.88	206.92
-1.46	0.23	198.98	200.34	171.24	172.23	208.51	207.40
-1.73	0.19	195.17	193.74	165.71	163.67	207.97	207.06
-2.04	0.15	184.28	184.52	158.98	158.28	207.70	204.09
-2.41	0.11	171.55	172.35	153.36	153.61	205.91	206.18
-0.60	0.38	200.23	197.58	200.54	200.36	204.06	202.03
-0.50	0.40	199.32	202.77	202.32	200.49	199.06	200.03
-0.41	0.42	193.73	192.48	205.64	202.56	198.94	199.40
-0.32	0.43	186.91	189.05	208.07	207.17	196.62	198.73
-0.22	0.45	181.68	182.41	211.03	209.77	192.99	194.56
-0.13	0.47	176.97	177.65	217.07	216.26	194.02	193.33
-0.04	0.49	168.58	166.95	221.04	219.15	192.40	190.89
0.05	0.51	162.88	159.23	221.40	220.15	187.15	192.39

Table B.11: Baseline ARL_1 results with corresponding $SDRL$, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, for IC ZINB Scenario 4 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
0.10	1.82	225.25	219.13	176.81
-0.01	1.64	249.31	301.31	254.66
-0.12	1.46	305.12	460.31	262.72
-0.26	1.28	370.56	488.15	288.52
-0.41	1.09	464.81	410.50	302.31
-0.59	0.91	714.36	322.44	297.69
-0.82	0.73	904.07	168.48	320.94
-1.10	0.55	1416.19	111.27	341.83
0.10	1.82	219.76	227.12	173.00
0.36	2.37	147.34	76.17	96.17
0.57	2.92	80.47	28.29	31.81
0.74	3.46	54.82	17.07	21.23
0.89	4.01	26.73	9.54	10.59
1.02	4.56	16.14	6.70	7.30
1.13	5.10	13.99	6.16	6.85
1.23	5.65	8.48	4.61	5.04
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
0.50	0.60	240.19	242.20	201.83
0.20	0.54	168.85	179.03	169.94
-0.09	0.48	124.78	152.84	192.60
-0.38	0.42	119.68	140.44	192.31
-0.69	0.36	98.85	124.98	154.81
-1.01	0.30	92.14	99.83	143.09
-1.37	0.24	69.55	94.01	130.90
-1.79	0.18	62.87	91.17	129.38
0.50	0.60	222.45	235.76	193.25
0.65	0.63	277.52	253.54	214.71
0.81	0.66	243.73	254.85	201.19
0.98	0.69	321.63	312.12	205.25
1.15	0.72	348.60	349.07	203.68
1.33	0.75	429.82	411.71	221.50
1.53	0.78	534.50	403.43	220.31
1.74	0.81	655.97	487.14	217.75

Table B.12: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, in IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
1.60	8.17	178.08	171.59	175.93
1.49	7.35	225.74	221.02	221.73
1.38	6.53	332.38	137.10	289.38
1.24	5.72	468.12	61.43	340.71
1.09	4.90	679.27	27.80	350.49
0.91	4.08	686.63	13.85	333.51
0.68	3.27	603.42	7.51	357.84
0.40	2.45	440.57	5.58	304.05
1.60	8.17	194.34	137.88	150.29
1.86	10.62	66.45	17.61	23.30
2.07	13.07	21.70	7.04	8.60
2.24	15.52	8.62	3.46	3.85
2.39	17.97	4.33	2.61	2.81
2.52	20.42	3.05	2.47	2.53
2.63	22.87	2.30	1.98	2.06
2.73	25.31	2.00	1.97	1.99
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
-0.60	0.38	213.49	185.00	157.21
-0.80	0.34	169.45	163.35	173.26
-1.00	0.30	166.02	126.57	165.79
-1.22	0.26	136.57	115.77	159.04
-1.46	0.23	124.11	134.59	153.20
-1.73	0.19	118.64	117.50	167.47
-2.04	0.15	103.58	132.57	178.84
-2.41	0.11	97.14	108.98	164.27
-0.60	0.38	202.97	167.51	175.83
-0.50	0.40	194.94	157.18	170.12
-0.41	0.42	181.50	177.53	175.84
-0.32	0.43	199.60	183.02	160.57
-0.22	0.45	269.77	196.43	173.66
-0.13	0.47	253.09	189.33	172.07
-0.04	0.49	235.93	171.23	155.42
0.05	0.51	230.32	193.79	134.00

Table B.13: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, in IC ZIP Scenario 2 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
0.10	1.82	139.58	169.94	193.91
-0.01	1.64	158.51	218.01	221.72
-0.12	1.46	236.57	299.95	244.65
-0.26	1.28	296.60	432.54	285.54
-0.41	1.09	364.99	539.48	292.77
-0.59	0.91	501.91	429.54	312.75
-0.82	0.73	769.87	357.89	308.73
-1.10	0.55	1109.11	196.10	305.79
0.10	1.82	157.44	169.02	192.88
0.36	2.37	74.94	66.16	94.99
0.57	2.92	44.60	31.39	42.26
0.74	3.46	31.32	19.05	27.54
0.89	4.01	19.93	11.92	16.28
1.02	4.56	13.63	9.86	12.18
1.13	5.10	10.72	6.61	9.11
1.23	5.65	6.79	5.24	6.34
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
0.50	0.60	142.35	148.67	179.81
0.20	0.54	122.78	134.53	180.63
-0.09	0.48	92.86	109.81	174.66
-0.38	0.42	82.08	97.69	136.78
-0.69	0.36	66.06	90.34	147.50
-1.01	0.30	51.05	72.68	128.29
-1.37	0.24	47.98	70.78	113.05
-1.79	0.18	41.95	54.42	102.37
0.50	0.60	140.06	160.33	185.94
0.65	0.63	182.24	195.43	208.84
0.81	0.66	193.01	191.28	163.86
0.98	0.69	207.81	215.93	201.68
1.15	0.72	263.87	266.31	208.93
1.33	0.75	282.52	308.15	168.37
1.53	0.78	338.90	363.38	207.28
1.74	0.81	440.32	420.49	198.81

Table B.14: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, in IC ZINB Scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
1.60	8.17	169.95	160.41	189.99
1.49	7.35	252.56	271.15	233.15
1.38	6.53	392.63	427.59	311.29
1.24	5.72	664.74	668.36	330.80
1.09	4.90	1006.90	793.48	356.00
0.91	4.08	1559.31	640.62	359.44
0.68	3.27	2272.57	340.51	320.25
0.40	2.45	3036.57	269.19	327.94
1.60	8.17	157.19	160.38	188.48
1.86	10.62	57.84	53.70	75.94
2.07	13.07	25.00	21.74	29.57
2.24	15.52	15.98	14.27	19.34
2.39	17.97	9.68	8.64	11.21
2.52	20.42	6.94	6.05	7.35
2.63	22.87	5.29	5.21	5.99
2.73	25.31	4.51	4.17	5.17
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
-0.60	0.38	154.35	160.16	182.23
-0.80	0.34	140.90	152.28	201.63
-1.00	0.30	131.81	147.18	172.96
-1.22	0.26	112.39	127.81	185.14
-1.46	0.23	103.72	117.27	182.01
-1.73	0.19	85.46	99.42	167.53
-2.04	0.15	90.42	99.63	181.83
-2.41	0.11	81.76	93.66	180.59
-0.60	0.38	138.79	141.04	201.33
-0.50	0.40	161.88	172.97	189.99
-0.41	0.42	166.18	168.31	177.31
-0.32	0.43	157.20	164.38	164.12
-0.22	0.45	206.31	179.52	184.74
-0.13	0.47	238.63	214.73	186.24
-0.04	0.49	191.73	187.16	162.11
0.05	0.51	246.62	232.97	214.32

Table B.15: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, L)-, (r^D, L)- and (r^Q, L)- Shewhart chart, in IC ZINB Scenario 4 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
0.10	1.82	281.56	214.37	207.58
-0.01	1.64	301.06	218.41	267.05
-0.12	1.46	288.76	148.56	340.44
-0.26	1.28	332.81	101.92	327.77
-0.41	1.09	389.38	64.69	338.09
-0.59	0.91	378.08	45.34	350.15
-0.82	0.73	307.22	34.34	352.23
-1.10	0.55	217.13	27.72	371.50
0.10	1.82	245.60	229.72	247.68
0.36	2.37	176.37	127.28	103.79
0.57	2.92	131.56	40.89	37.81
0.74	3.46	91.44	20.75	19.89
0.89	4.01	49.17	12.51	12.53
1.02	4.56	31.66	9.65	9.72
1.13	5.10	17.62	6.89	7.03
1.23	5.65	14.74	6.21	6.33
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
0.50	0.60	249.83	243.04	205.87
0.20	0.54	248.75	216.14	212.35
-0.09	0.48	243.69	185.77	214.90
-0.38	0.42	222.11	182.55	216.40
-0.69	0.36	189.00	142.10	180.02
-1.01	0.30	189.28	174.84	214.97
-1.37	0.24	130.45	150.88	163.68
-1.79	0.18	133.86	126.94	148.35
0.50	0.60	276.44	244.50	226.66
0.65	0.63	231.19	239.43	201.97
0.81	0.66	223.40	291.33	237.74
0.98	0.69	253.47	281.97	204.91
1.15	0.72	204.12	311.13	226.67
1.33	0.75	195.59	321.01	259.62
1.53	0.78	181.31	350.76	219.19
1.74	0.81	188.88	404.57	251.29

Table B.16: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, in IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
1.60	8.17	227.79	203.27	198.18
1.49	7.35	272.13	133.81	282.44
1.38	6.53	286.30	55.49	348.74
1.24	5.72	371.17	24.81	381.49
1.09	4.90	330.92	13.63	403.74
0.91	4.08	304.96	7.89	358.32
0.68	3.27	228.03	5.99	376.24
0.40	2.45	126.50	4.05	327.11
1.60	8.17	257.54	209.63	233.29
1.86	10.62	126.58	34.85	31.64
2.07	13.07	55.92	8.95	8.97
2.24	15.52	15.27	4.38	4.28
2.39	17.97	5.95	3.02	2.94
2.52	20.42	3.79	2.33	2.27
2.63	22.87	2.74	2.27	2.26
2.73	25.31	2.31	2.12	2.16
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
-0.60	0.38	228.64	206.82	225.01
-0.80	0.34	252.40	202.03	227.41
-1.00	0.30	276.91	186.09	235.02
-1.22	0.26	253.34	172.95	223.54
-1.46	0.23	239.42	155.17	231.94
-1.73	0.19	213.69	157.84	257.92
-2.04	0.15	223.12	164.05	261.71
-2.41	0.11	173.88	151.40	243.34
-0.60	0.38	229.61	208.82	248.38
-0.50	0.40	221.02	209.34	212.58
-0.41	0.42	202.41	220.22	217.42
-0.32	0.43	246.00	197.63	208.03
-0.22	0.45	241.36	238.39	220.34
-0.13	0.47	197.96	220.82	196.90
-0.04	0.49	201.81	242.30	206.76
0.05	0.51	202.78	278.56	217.56

Table B.17: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, in IC ZIP Scenario 2 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
0.10	1.82	200.38	223.81	224.49
-0.01	1.64	211.26	235.00	264.49
-0.12	1.46	242.34	218.16	267.42
-0.26	1.28	264.43	140.69	324.63
-0.41	1.09	305.42	115.14	324.64
-0.59	0.91	275.84	87.69	342.50
-0.82	0.73	250.88	60.98	372.05
-1.10	0.55	220.78	47.44	343.60
0.10	1.82	233.81	244.62	231.59
0.36	2.37	136.94	115.02	106.89
0.57	2.92	110.72	70.44	58.84
0.74	3.46	72.41	31.86	28.82
0.89	4.01	38.97	19.82	19.71
1.02	4.56	28.91	13.06	12.81
1.13	5.10	19.33	8.95	8.27
1.23	5.65	15.05	7.91	7.57
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
0.50	0.60	210.74	212.29	236.00
0.20	0.54	211.19	223.35	208.72
-0.09	0.48	206.00	185.74	184.62
-0.38	0.42	204.76	175.91	185.95
-0.69	0.36	169.34	154.27	163.31
-1.01	0.30	148.41	171.55	179.38
-1.37	0.24	137.41	143.57	137.98
-1.79	0.18	103.36	119.20	118.97
0.50	0.60	199.22	235.25	225.81
0.65	0.63	221.63	246.37	220.18
0.81	0.66	219.69	240.54	208.59
0.98	0.69	192.12	262.45	212.35
1.15	0.72	181.00	256.23	251.97
1.33	0.75	196.16	287.20	263.38
1.53	0.78	169.91	300.35	231.50
1.74	0.81	153.56	305.31	248.62

Table B.18: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, in IC ZINB Scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

		OC scenario due to decreased/increased $E[\lambda^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
β_0^{OC}	$E[\lambda^{OC}]$	ARL_1	ARL_1	ARL_1
1.60	8.17	195.18	210.11	183.53
1.49	7.35	264.92	207.24	220.04
1.38	6.53	332.84	221.27	275.48
1.24	5.72	398.18	185.28	306.54
1.09	4.90	377.02	149.81	368.96
0.91	4.08	373.75	93.65	336.06
0.68	3.27	322.15	63.51	354.00
0.40	2.45	272.88	41.64	325.71
1.60	8.17	198.84	163.54	177.75
1.86	10.62	81.25	70.11	66.83
2.07	13.07	35.64	31.82	30.35
2.24	15.52	21.39	19.83	18.56
2.39	17.97	11.11	9.77	9.56
2.52	20.42	8.90	7.96	7.63
2.63	22.87	6.08	5.04	5.14
2.73	25.31	5.78	5.26	5.26
		OC scenario due to decreased/increased $E[p^{OC}]$		
		(r^P, L)-Shewhart	(r^D, L)-Shewhart	(r^Q, L)-Shewhart
γ_0^{OC}	$E[p^{OC}]$	ARL_1	ARL_1	ARL_1
-0.60	0.38	203.34	200.29	189.34
-0.80	0.34	208.04	173.71	180.22
-1.00	0.30	207.93	183.53	208.15
-1.22	0.26	202.56	170.00	209.29
-1.46	0.23	217.50	178.79	184.01
-1.73	0.19	203.69	154.63	206.86
-2.04	0.15	184.25	149.53	177.68
-2.41	0.11	174.84	146.79	184.46
-0.60	0.38	223.96	204.56	172.38
-0.50	0.40	237.65	180.47	188.84
-0.41	0.42	183.86	191.81	167.53
-0.32	0.43	213.79	217.37	198.80
-0.22	0.45	208.19	207.88	172.23
-0.13	0.47	170.99	186.10	178.12
-0.04	0.49	192.90	240.34	182.18
0.05	0.51	165.09	213.50	157.78

Table B.19: ARL_1 results with corresponding $SDRL$ while taking into account Phase I estimation effects, for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)- Shewhart chart, in IC ZINB Scenario 4 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

B.6 Additional results of the Gamma GLM-based TBE charts

				No OC signal (% of total)		
IC scenario	Residuals	ARL_0	$SDRL_0$	$n = 2000$	$n = 3000$	$n = 4000$
ZIP 1	Pearson	199.31	194.77	0.00	0.00	0.00
ZIP 1	Deviance	199.63	195.01	0.00	0.00	0.00
ZIP 1	Quantile	199.80	195.22	0.00	0.00	0.00
ZIP 2	Pearson	202.09	202.17	0.00	0.00	0.00
ZIP 2	Deviance	201.43	201.73	0.00	0.00	0.00
ZIP 2	Quantile	201.74	201.90	0.00	0.00	0.00
ZINB 3	Pearson	201.35	199.71	0.01	0.00	0.00
ZINB 3	Deviance	200.86	199.30	0.01	0.00	0.00
ZINB 3	Quantile	200.88	199.27	0.01	0.00	0.00
ZINB 4	Pearson	201.49	200.42	0.00	0.00	0.00
ZINB 4	Deviance	201.56	200.73	0.00	0.00	0.00
ZINB 4	Quantile	201.54	201.29	0.00	0.00	0.00

Table B.20: Percentage of runs with no OC signal for the Gamma GLM-based TBE chart, for $n = 2000, 3000, 4000$.

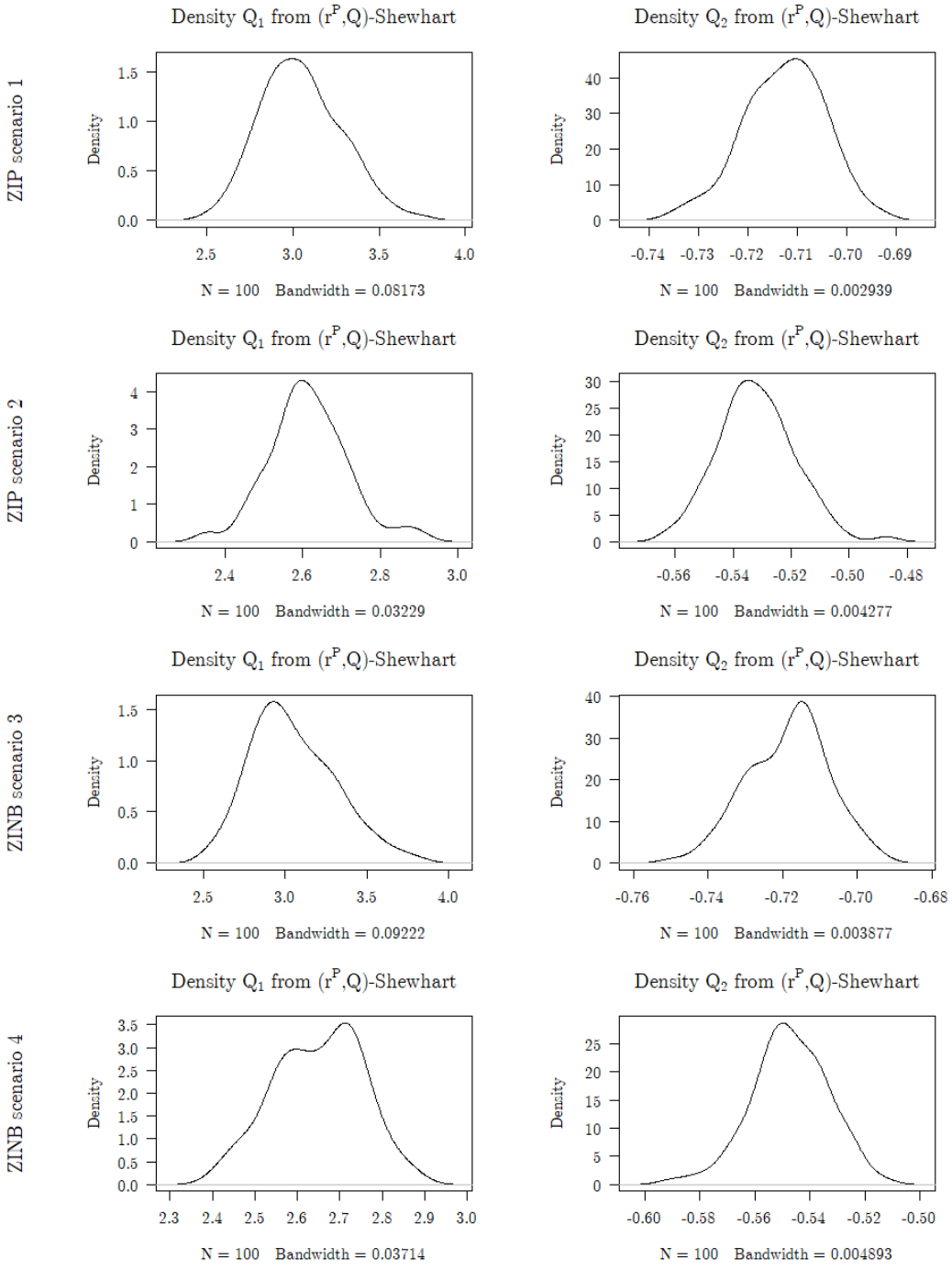

Figure B.24: The density of the charting constants Q_1 and Q_2, solved 100 times for performance evaluation of the (r^P, Q)-TBE chart in each IC scenario.

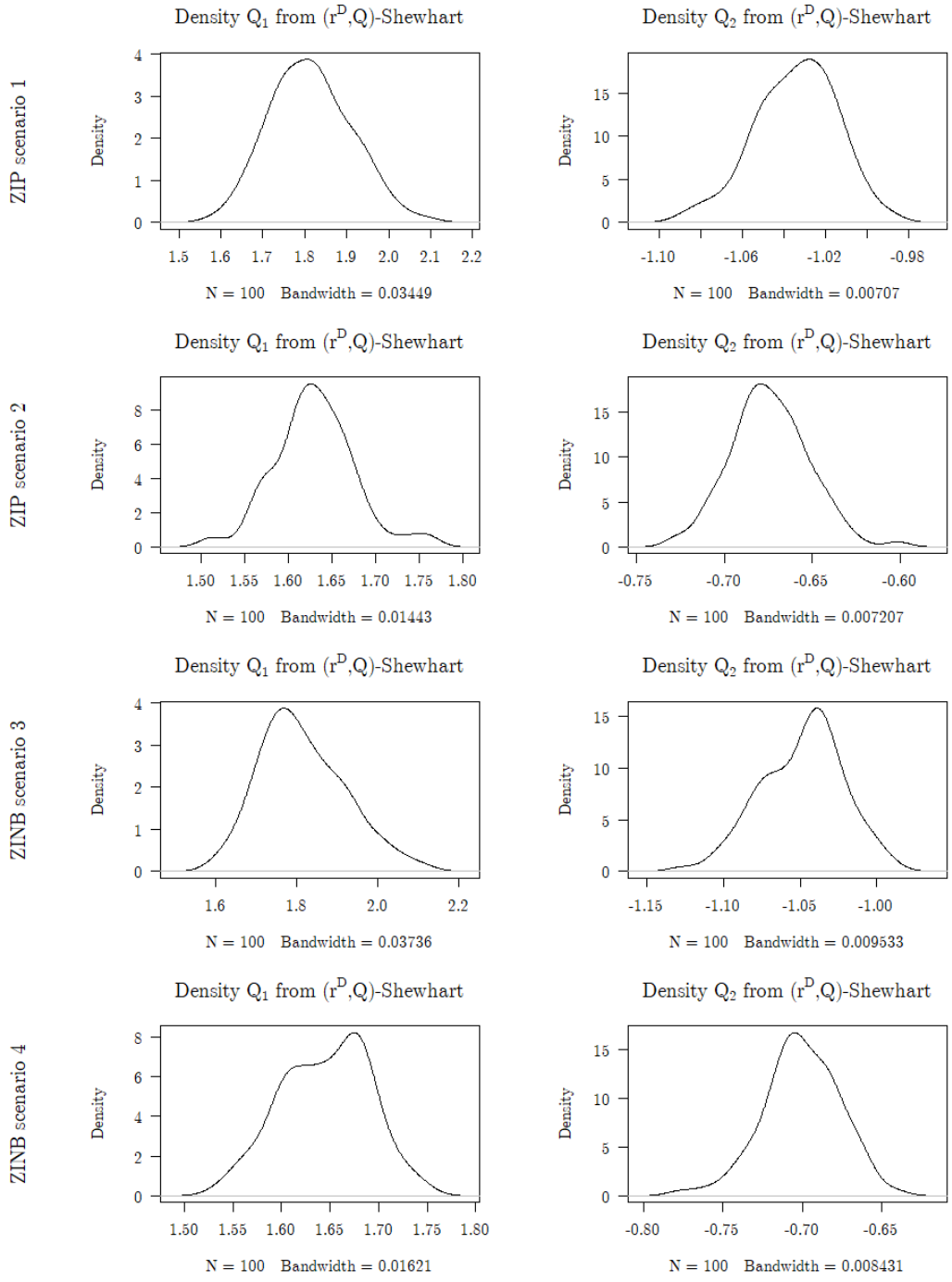

Figure B.25: The density of the charting constants Q_1 and Q_2, solved 100 times for performance evaluation of the (r^D, Q)-TBE chart in each IC scenario.

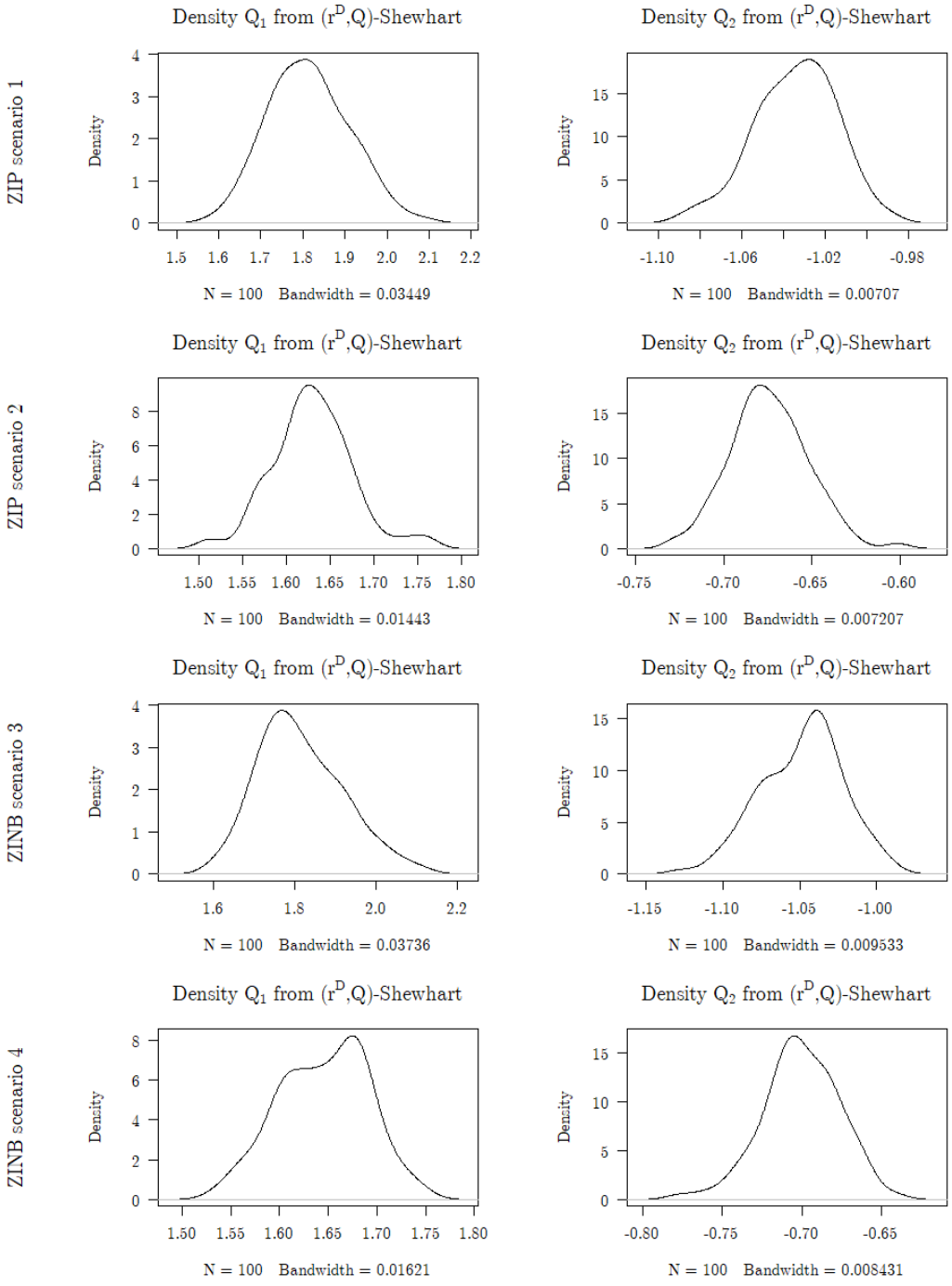

Figure B.26: The density of the charting constants Q_1 and Q_2, solved 100 times for performance evaluation of the (r^Q, Q)-TBE chart in each IC scenario.

OC scenario due to decreased/increased $E[\lambda^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
β_0^{OC}	$E[\lambda^{OC}]$	ALI_1	ALI_1	ALI_1
0.10	1.82	210.50	207.59	209.44
-0.01	1.64	217.63	217.30	217.02
-0.12	1.46	204.18	203.62	204.00
-0.26	1.28	179.30	178.94	179.49
-0.41	1.09	174.31	172.51	174.20
-0.59	0.91	140.00	140.08	139.93
-0.82	0.73	105.28	104.27	105.26
-1.10	0.55	77.36	76.53	77.17
0.10	1.82	221.18	218.98	220.98
0.36	2.37	195.99	192.39	195.70
0.57	2.92	203.02	201.73	201.65
0.74	3.46	181.44	180.79	181.38
0.89	4.01	158.26	156.91	157.68
1.02	4.56	146.51	146.11	146.58
1.13	5.10	145.20	144.25	145.30
1.23	5.65	133.86	131.13	132.30
OC scenario due to decreased/increased $E[p^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
γ_0^{OC}	$E[p^{OC}]$	ALI_1	ALI_1	ALI_1
0.50	0.60	206.44	204.75	205.71
0.20	0.54	225.05	223.00	224.96
-0.09	0.48	196.07	195.72	196.62
-0.38	0.42	145.04	144.80	145.19
-0.69	0.36	116.05	113.87	115.99
-1.01	0.30	91.37	90.16	91.80
-1.37	0.24	67.50	67.49	67.64
-1.79	0.18	55.06	54.85	54.91
0.50	0.60	211.36	210.69	211.17
0.65	0.63	198.13	198.14	197.86
0.81	0.66	171.33	169.56	171.22
0.98	0.69	137.11	135.91	137.00
1.15	0.72	102.03	101.99	101.98
1.33	0.75	84.14	84.11	84.13
1.53	0.78	64.88	65.04	64.84
1.74	0.81	53.98	53.96	53.96

Table B.21: ALI_1 results with corresponding $SDLI$ for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart, in IC ZIP Scenario 1 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

OC scenario due to decreased/increased $E[\lambda^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
β_0^{OC}	$E[\lambda^{OC}]$	ALI_1	ALI_1	ALI_1
1.60	8.17	196.19	196.24	196.44
1.49	7.35	196.18	196.13	195.13
1.38	6.53	203.17	203.17	203.83
1.24	5.72	217.58	217.36	215.73
1.09	4.90	203.16	203.01	204.07
0.91	4.08	200.21	200.53	200.74
0.68	3.27	184.36	184.62	184.76
0.40	2.45	151.26	151.18	151.67
1.60	8.17	187.21	187.15	187.47
1.86	10.62	178.48	177.25	177.80
2.07	13.07	168.80	169.15	169.64
2.24	15.52	169.23	168.92	169.25
2.39	17.97	153.45	153.32	153.62
2.52	20.42	145.68	145.42	145.39
2.63	22.87	141.41	141.18	141.83
2.73	25.31	146.86	146.80	146.85
OC scenario due to decreased/increased $E[p^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
γ_0^{OC}	$E[p^{OC}]$	ALI_1	ALI_1	ALI_1
-0.60	0.38	207.10	207.01	207.47
-0.80	0.34	210.89	210.32	211.48
-1.00	0.30	207.22	205.24	205.06
-1.22	0.26	190.89	190.73	191.53
-1.46	0.23	164.37	163.04	163.62
-1.73	0.19	141.76	141.70	141.69
-2.04	0.15	115.67	113.99	113.94
-2.41	0.11	85.26	85.14	85.36
-0.60	0.38	201.02	201.06	202.02
-0.50	0.40	168.09	168.04	168.83
-0.41	0.42	170.82	170.64	170.65
-0.32	0.43	155.09	155.05	155.49
-0.22	0.45	118.78	118.47	118.50
-0.13	0.47	119.10	119.06	119.79
-0.04	0.49	106.06	106.03	105.88
0.05	0.51	89.19	89.34	89.46

Table B.22: ALI_1 results with corresponding $SDLI$ for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart, in IC ZIP Scenario 2 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

OC scenario due to decreased/increased $E[\lambda^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
β_0^{OC}	$E[\lambda^{OC}]$	ALI_1	ALI_1	ALI_1
0.10	1.82	193.88	197.67	193.46
-0.01	1.64	178.74	179.56	178.80
-0.12	1.46	180.28	183.04	180.54
-0.26	1.28	177.47	180.44	177.42
-0.41	1.09	152.99	156.61	153.76
-0.59	0.91	130.43	130.28	129.49
-0.82	0.73	102.57	103.67	102.44
-1.10	0.55	77.71	78.53	77.67
0.10	1.82	203.64	206.09	202.63
0.36	2.37	190.27	194.67	193.10
0.57	2.92	173.67	177.14	175.40
0.74	3.46	167.40	168.73	166.54
0.89	4.01	140.15	141.75	140.71
1.02	4.56	138.06	138.64	138.45
1.13	5.10	127.93	129.61	127.85
1.23	5.65	120.55	121.26	120.56
OC scenario due to decreased/increased $E[p^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
γ_0^{OC}	$E[p^{OC}]$	ALI_1	ALI_1	ALI_1
0.50	0.60	186.05	187.74	185.47
0.20	0.54	201.13	205.18	201.57
-0.09	0.48	161.71	167.04	161.09
-0.38	0.42	129.59	130.24	129.18
-0.69	0.36	106.60	107.74	106.34
-1.01	0.30	84.58	84.86	84.34
-1.37	0.24	64.81	65.97	65.07
-1.79	0.18	55.81	56.09	55.59
0.50	0.60	190.34	193.40	190.35
0.65	0.63	186.48	188.77	186.30
0.81	0.66	145.45	146.96	145.31
0.98	0.69	131.69	132.64	131.61
1.15	0.72	99.84	100.87	99.86
1.33	0.75	83.87	84.36	83.84
1.53	0.78	65.47	66.64	65.57
1.74	0.81	55.00	55.39	55.05

Table B.23: ALI_1 results with corresponding $SDLI$ for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart, in IC ZINB Scenario 3 with $E[p^{IC}] = 0.60$ and $E[\lambda^{IC}] = 1.82$.

OC scenario due to decreased/increased $E[\lambda^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
β_0^{OC}	$E[\lambda^{OC}]$	ALI_1	ALI_1	ALI_1
1.60	8.17	199.88	202.79	202.50
1.49	7.35	184.16	186.58	186.01
1.38	6.53	183.97	186.98	186.69
1.24	5.72	175.22	176.97	176.74
1.09	4.90	181.66	184.94	184.86
0.91	4.08	164.84	166.79	166.72
0.68	3.27	142.25	143.22	143.15
0.40	2.45	106.78	107.16	107.10
1.60	8.17	186.11	189.80	190.73
1.86	10.62	175.51	178.57	178.35
2.07	13.07	161.95	163.18	163.28
2.24	15.52	168.48	170.15	170.09
2.39	17.97	149.88	152.12	151.95
2.52	20.42	137.32	138.41	138.30
2.63	22.87	152.16	152.57	152.60
2.73	25.31	138.33	141.79	141.85
OC scenario due to decreased/increased $E[p^{OC}]$				
		(r^P, L)-TBE	(r^D, L)-TBE	(r^Q, L)-TBE
γ_0^{OC}	$E[p^{OC}]$	ALI_1	ALI_1	ALI_1
-0.60	0.38	183.81	186.35	186.15
-0.80	0.34	185.22	188.70	188.08
-1.00	0.30	191.82	195.65	195.96
-1.22	0.26	172.82	174.10	175.06
-1.46	0.23	157.01	158.70	158.83
-1.73	0.19	144.29	144.94	144.87
-2.04	0.15	98.58	99.72	99.60
-2.41	0.11	86.26	88.06	88.22
-0.60	0.38	182.58	186.03	185.28
-0.50	0.40	162.52	164.21	164.15
-0.41	0.42	151.89	156.36	156.55
-0.32	0.43	149.54	151.59	151.43
-0.22	0.45	120.33	121.18	121.09
-0.13	0.47	108.60	110.41	110.51
-0.04	0.49	100.82	101.97	102.00
0.05	0.51	79.84	80.64	80.42

Table B.24: ALI_1 results with corresponding $SDLI$ for the (r^P, Q)-, (r^D, Q)- and (r^Q, Q)-TBE chart, in IC ZINB Scenario 4 with $E[p^{IC}] = 0.38$ and $E[\lambda^{IC}] = 8.17$.

C | Appendix: R code

This appendix contains a fraction of the R code that was used to implement all simulations, of which the most important functions are provided. Section B.1 contains the code that was executed to run the simulations of the ZIP-EWMA chart. Section C.2 contains the most important functions to execute a baseline performance analysis of the ZIP and ZINB regression based Shewhart chart with Pearson, deviance and randomised quantile residuals. An example file that shows how the baseline performance is executed is provided in Section C.3. Similarly, the functions of the Gamma GLM-based TBE charts are provided in Section C.4, after which an example file for TBE simulation execution is provided in Section C.5.

C.1 Simulations for the ZIP-EWMA chart

```
library(mc2d)
library(VGAM)
library(xtable)
library(RColorBrewer)

#Defining run length function
RL <- function (x,u,l,n) {min(min(n+1,which(x > u)), min(n+1,which(x < l))) - 1}

#Defining ARL function for ZIP EWMA chart
Simulate_ARL <- function (L, p, lambda, w, N, n) {
  Y <- matrix(rzipois(n*N, lambda, p),N,n)
  V <- t(apply(Y, 1, Reduce, f = function (v,y) w*y+(1-w)*v, init = (1-p)*lambda, accumulate = TRUE))
  Bandw <- sapply(0:n+1, function (i) L*sqrt((w/(2-w))*(1-(1-w)^(2*i))*((1-p)*(lambda+p*(lambda^2)))))
  UCL <- (1-p)*lambda + Bandw
  LCL <- (1-p)*lambda - Bandw
  ARL <- mean(apply(V,1,RL, u=UCL, l=LCL, n=n))
}

#Solving for L1, for a fixed ARL value
ineq <- function(L, P, Lambda, W, N, n, SET_ARL) {Simulate_ARL(L, P, Lambda, W, N, n) - SET_ARL}
Solve_L <- function(p, lambda, w, N1, n1, set_ARL, Search_Range) {uniroot(ineq, Search_Range, P = p, Lambda
      = lambda, W = w, N = N1, n = n1, SET_ARL=set_ARL, tol=0.001)$root}


#Searching for a proper simulation size n
#First: Solving L for multiple values of ARL and w and MLE parameters p and lambda
i <- 0
j <- 0
lambda <- 3
w <- 0.2
P <- c(0.3,0.5,0.8)
ARL <- c(200, 370, 500)
results_gues_L <- matrix(rep(9*6,0),9,6)
try_L <- seq(2,3.5,0.2)
```

```
ans <- rep(0,length(try_L))
for (p in P) {
  for (arl in ARL) {

        #Quick solve to determine search range
        for (l in try_L){
          i <- i + 1
          ans[i] <- Simulate_ARL(l ,p, lambda, w, 1000, 5000)
          if (i > 1) {
            if (ans[i-1] < arl & ans[i] > arl) {
              Search_Range <- c(l-0.3,l+0.1)
              i <- 0
              break
            }}}

        #Solving for L and storing results
        j <- j + 1
        print(j)
        L <- Solve_L(p, lambda, w, 1000, 5000, arl, Search_Range)
        results_gues_L[j,] <- c(p,lambda, w, arl, round(L,4), L)
}}

write.csv(results_gues_L ,"Guess_for_L.csv", row.names = FALSE, quote = FALSE)

#Second: Calculating the amount of runs with no OC signals for ZIP-EWMA with guessed L values
df <- read.csv("Guess_for_L.csv")
L_values <- df[,6]
p_values <- df[,1]
arl_values <- df[,4]
m_values <- c(2,5,10)
N <- 10000
lambda <- 3
w <- 0.2
results_n <- matrix(rep(0,8*length(L_values)*length(m_values)), length(L_values)*length(m_values), 8)
j <- 0

for (i in 1:length(L_values)) {
  L <- L_values[i]
  p <- p_values[i]
  arl <- arl_values[i]
  for (m in m_values) {
    j <- j + 1
    n <- m*arl
    Y <- matrix(rzipois(n*N, lambda, p),N,n)
    V <- t(apply(Y, 1, Reduce, f = function (v,y) w*y+(1-w)*v, init = (1-p)*lambda, accumulate = TRUE))
    Bandw <- sapply(0:n+1, function (i) L*sqrt((w/(2-w))*(1-(1-w)^(2*i))*((1-p)*(lambda+p*(lambda^2)))))
    UCL <- (1-p)*lambda + Bandw
    LCL <- (1-p)*lambda - Bandw
    RLs <- apply(V,1,RL, u=UCL, l=LCL, n=n)
    ARL <- mean(RLs)
    SDRL <- sd(RLs)
    results_n[j,] <- c(p, arl, L, n, mean(RLs), sd(RLs), sum(RLs == n), sum(RLs == n)/100)
    print(results_n)
  }
}

results_n
write.csv(results_n ,"Results_n.csv", row.names = FALSE, quote = FALSE)

#Solving L for multiple values of ARL and w and MLE parameters p and lambda
i <- 0
```

```
j <- 0
P <- c(0.3, 0.5, 0.8)
LAMBDA <- c(3, 4)
weights <- c(0.2, 0.3)
ARL <- c(200, 370, 500)
results <- matrix(rep(36*6,0),36,6)
try_L <- seq(2,4.5,0.2)
ans <- rep(0,length(try_L))
for (p in P) {
  for (lambda in LAMBDA) {
    for (w in weights) {
      for (arl in ARL) {

      #Quick solve to determine search range
      for (l in try_L){
        i <- i + 1
        ans[i] <- Simulate_ARL(l ,p, lambda, w, N = 1000, n = 10*arl)
        if (i > 1) {
          if (ans[i-1] < arl & ans[i] > arl) {
            Search_Range <- c(l-0.3,l+0.1)
            i <- 0
            break
          }}}

      #Solving for L and storing results
      j <- j + 1
      print(j)
      L <- Solve_L(p, lambda, w, N = 10000, n = 10*arl, arl, Search_Range)
      results[j,] <- c(p,lambda, w, arl, round(L,4), L)
}}}}

results
write.csv(results ,"Results_L.csv", row.names = FALSE, quote = FALSE)
write.csv(results[,5] ,"Results_L_rounded.csv", row.names = FALSE, quote = FALSE)


#Defining function to Calculate ARL1
Simulate_ARL1 <- function (L, p, p1, lambda, lambda1, w, N, n) {
  Y <- matrix(rzipois(n*N, lambda1, p1),N,n)
  V <- t(apply(Y, 1, Reduce, f = function (v,y) w*y+(1-w)*v, init = (1-p)*lambda, accumulate = TRUE))
  Bandw <- sapply(0:n+1, function (i) L*sqrt((w/(2-w))*(1-(1-w)^(2*i))*((1-p)*(lambda+p*(lambda^2)))))
  UCL <- (1-p)*lambda + Bandw
  LCL <- (1-p)*lambda - Bandw
  ARL <- mean(apply(V,1,RL, u=UCL, l=LCL, n=n))
  return(ARL)
}

#Initializing ARL1 calculations for various OC scenarios
DF <- read.csv("Results_L.csv")
idx <- c(1,2,3,13,14,15,25,26,27)
lambda <- 3
w <- 0.2
p_values <- results[idx,1]
ARL0_values <- results[idx,4]
L_values <- results[idx,6]
delta_p <- c(-0.1, -0.2, -0.3)
delta_lambda <- c(0.5, 1, 2)
OC_results <- matrix(rep(9*9*3,0),9,9*3)

#Calculating ARL1 for IC(p, lambda, w) = (0.3, 3, 0.2), (0.5, 3, 0.2) and (0.8, 3, 0.2)
j <- 0
```

```
z <- 3
for (i in 1:length(L_values)) {
    L <- L_values[i]
    ARL0 <- ARL0_values[i]
    p <- p_values[i]
for (Dp in delta_p) {
    j <- j + 1
    print(j)
    OC_results[j,1+3*z] <- p + Dp
    OC_results[j,2+3*z] <- lambda
    OC_results[j,3+3*z] <- Simulate_ARL1(L, p, p + Dp, lambda, lambda, w, N = 10000, n = 10*ARL0)}
for (Dl in delta_lambda) {
    j <- j + 1
    print(j)
    OC_results[j,1+3*z] <- p
    OC_results[j,2+3*z] <- lambda + Dl
    OC_results[j,3+3*z] <- Simulate_ARL1(L, p, p, lambda, lambda + Dl, w, N = 10000, n = 10*ARL0)}
for (k in 1:3) {
  j <- j + 1
  print(j)
  OC_results[j,1+3*z] <- p + delta_p[k]
  OC_results[j,2+3*z] <- lambda + delta_lambda[k]
  OC_results[j,3+3*z] <- Simulate_ARL1(L, p, p + delta_p[k], lambda, lambda + delta_lambda[k], w, 10000, n
       = 10*ARL0)}
z <- z + 1
j <- 0}
```

C.2 Functions for baseline performance evaluation

```
library(VGAM)
library(statip)
library(pscl)

#Simulating ZIP data
Simulate.ZIP.Data = function (b0, b1, g0, g1, n, N, muX, varX) {
  X = matrix(rnorm(n*N, muX, sqrt(varX)),N,n)
  p = exp(g0+g1*X)/(1+exp(g0+g1*X))
  lamb = exp(b0 + b1*X)
  c = apply(p, 1:2, rbinom, n=1, size=1) ;c
  Y = matrix(0,N,n)
  Y[c==0] = rpois(sum(c==0), lamb[c==0])
  return(list(Y,X))
}

#Simulating ZINB data
Simulate.ZINB.Data = function (b0, b1, g0, g1, tau, n, N, muX, varX) {
  X = matrix(rnorm(n*N, muX, sqrt(varX)),N,n)
  p = exp(g0+g1*X)/(1+exp(g0+g1*X))
  lamb = exp(b0 + b1*X)
  c = apply(p, 1:2, rbinom, n=1, size=1)
  Y = matrix(0,N,n)
  Y[c==0] = rnbinom(sum(c==0), size=tau, mu=lamb[c==0])
  return(list(Y,X))
}

#Pearson residuals function ZIP, ZINB
pearson.residuals = function (data, fixed.model, dist) {
  n = length(data)/2
  Y = data[1:n]
```

```
  X = data[(n+1):(2*n)]
  b0 = fixed.model[1]
  b1 = fixed.model[2]
  g0 = fixed.model[3]
  g1 = fixed.model[4]
  p = sapply(X, function (X, g0, g1) exp(g0+g1*X)/(1+exp(g0+g1*X)), g0 = g0, g1 = g1)
  lamb = sapply(X, function (X, b0, b1) exp(b0 + b1*X), b0 = b0, b1 = b1)
  if (dist == "ZIP") {
    rp = (Y - (1-p)*lamb)/sqrt((1-p)*(lamb+p*(lamb^2)))
    return(rp) }
  if (dist == "ZINB") {
    tau = fixed.model[7]
    rp = (Y - (1-p)*lamb)/sqrt(lamb*(1-p)*(1+(p*lamb)+(lamb/tau)))
    return(rp) }
}

#Deviance residuals function ZIP, ZINB
deviance.residuals = function (data, fixed.model, Y, X, dist) {
  n = length(data)/2
  Y = data[1:n]
  X = data[(n+1):(2*n)]
  b0 = fixed.model[1]
  b1 = fixed.model[2]
  g0 = fixed.model[3]
  g1 = fixed.model[4]
  p = sapply(X, function (X, g0, g1) exp(g0+g1*X)/(1+exp(g0+g1*X)), g0 = g0, g1 = g1)
  lamb = sapply(X, function (X, b0, b1) exp(b0 + b1*X), b0 = b0, b1 = b1)
  predictions = (1-p)*lamb
  r = Y - predictions
  if (dist == "ZIP") {
    logLik.Pois = log(dpois(Y,Y))
    logLik.ZIP = log(dzipois(Y, lambda = lamb, pstr0 = p))
    rd = sign(r)*sqrt(2*(logLik.Pois - logLik.ZIP))
    return(rd)
  }
  if (dist == "ZINB") {
    tau = fixed.model[7]
    logLik.NB = log(dnbinom(Y, size = tau, mu = Y))
    logLik.ZINB = log(dzinegbin(Y, size = tau, munb = lamb, pstr0 = p))
    rd = sign(r)*sqrt(2*(logLik.NB - logLik.ZINB))
    return(rd)
  }
}

#Function for randomized quantile residuals ZIP, ZINB
RQ.residuals = function (data, fixed.model, Y, X, dist) {
  n = length(data)/2
  Y = data[1:n]
  X = data[(n+1):(2*n)]
  b0 = fixed.model[1]
  b1 = fixed.model[2]
  g0 = fixed.model[3]
  g1 = fixed.model[4]
  p = sapply(X, function (X, g0, g1) exp(g0+g1*X)/(1+exp(g0+g1*X)), g0 = g0, g1 = g1)
  lamb = sapply(X, function (X, b0, b1) exp(b0 + b1*X), b0 = b0, b1 = b1)

  if (dist == "ZINB") {
    tau = fixed.model[7]
    a = pzinegbin(Y-1, size = tau, pstr0 = p, munb = lamb)
    b = pzinegbin(Y,size = tau, pstr0 = p, munb = lamb)
    u = runif(length(a), a, b)
```

```
    rq = qnorm(u)
    return(rq)
  }
  if (dist == "ZIP") {
    a = pzipois(Y-1, lambda = lamb, pstr0 = p)
    b = pzipois(Y, lambda = lamb, pstr0 =p)
    u = runif(length(a), a, b)
    rq = qnorm(u)
    return(rq)
  }
}

#Function for raw quantile residuals ZIP, ZINB
raw.residuals = function (data, fixed.model, Y, X, dist) {
  n = length(data)/2
  Y = data[1:n]
  X = data[(n+1):(2*n)]
  b0 = fixed.model[1]
  b1 = fixed.model[2]
  g0 = fixed.model[3]
  g1 = fixed.model[4]
  p = sapply(X, function (X, g0, g1) exp(g0+g1*X)/(1+exp(g0+g1*X)), g0 = g0, g1 = g1)
  lamb = sapply(X, function (X, b0, b1) exp(b0 + b1*X), b0 = b0, b1 = b1)
  predictions = (1-p)*lamb
  r = Y - predictions
}

#Run length function
RL <- function (x,u,l,n) {min(min(n, which(x > u)), min(n, which(x < l)))}

#Shewharts ARL0 simulation
Shewhart.ARL0 <- function(r0, mu.r, sd.r, L) {
  n = ncol(r0)
  N = nrow(r0)
  UCL <- mu.r + L*sd.r
  LCL <- mu.r - L*sd.r
  RLs <- apply(r0, 1, RL, u=UCL, l=LCL, n=n)
  #if (sum(RLs == n) > 0.001*N) {warning("Simulation size too small, in terms of n = nr. of columns") ;
      print(sum(RLs == n))}
  ARL0 <- mean(RLs)
  return(ARL0)
}

#Inequality to solve for L: to be solved by uniroot
inequality = function(L, r0, mu.r, sd.r, SET.ARL) {Shewhart.ARL0(r0 = r0, mu.r=mu.r, sd.r=sd.r, L = L) -
    SET.ARL}

#Solving for charting constant L
Shewhart.Solve.L = function(r0, mu.r, sd.r, set.ARL, SR.N = 1000, lowerBound.SR.L = 1.5, upperBound.SR.L =
    5.5, stepsize.try.L = 0.2, safety.margin.SR = 0.1) {

  #Determining simulation size
  n = ncol(r0)
  N = nrow(r0)

  #Quick solve to determine search range
  try.L = seq(lowerBound.SR.L, upperBound.SR.L, stepsize.try.L)
  ans2 = 0
  i = 0
  l = 0
  for (l in try.L){
```

```
    i = i+1
    ans1 = Shewhart.ARL0(r0[1:min(N,SR.N),], mu.r, sd.r, l)
    if (ans2 < set.ARL & set.ARL < ans1) {
      if (i == 1) {warning("lowerBound.SR.L too high"); break}
      search.range = c(l - stepsize.try.L - safety.margin.SR, l + safety.margin.SR)
      break
    }
    if (i == length(try.L)) {warning("upperBound.SR.L too low"); break}
    ans2 = ans1
  }

  #Solving for L
  Solved.L = NA
  Solved.L = uniroot(inequality, search.range, r0 = r0, mu.r=mu.r, sd.r=sd.r, SET.ARL=set.ARL, tol=0.001)$
      root
  if (is.na(Solved.L)==TRUE) {warning("safety.margin.SR is probably too small"); break}
  return(Solved.L)
}

#Function to calculate average number of proportion r > q1
Calculate.Alpha1 = function(resid, Q1, n){
  above.Q1 = apply(resid, 1, function(resid,Q1){sum(resid > Q1)}, Q1=Q1)
  proportion.above = above.Q1/n
  alpha1 = mean(proportion.above)
  return(alpha1)
}

#Function to calculate average number of proportion r < q2
Calculate.Alpha2 = function(resid, Q2, n){
  below.Q2 = apply(resid, 1, function(resid,Q2){sum(resid < Q2)}, Q2=Q2)
  proportion.below = below.Q2/n
  alpha2 = mean(proportion.below)
  return(alpha2)
}

#Inequalities to solve probability control limits, for a given quantile level
inequality.Q1 = function(resid, Q1, n, SET.Alpha1) {Calculate.Alpha1(resid = resid, Q1 = Q1, n = n) - SET.
    Alpha1}
inequality.Q2 = function(resid, Q2, n, SET.Alpha2) {Calculate.Alpha2(resid = resid, Q2 = Q2, n = n) - SET.
    Alpha2}

Solve.Q = function(resid, set.Alpha){
  #Determining simulation size
  n = ncol(resid)
  N = nrow(resid)
  #Solving Q1 and Q2
  search.range = c(min(resid), max(resid))
  Solved.Q1 = NA
  Solved.Q1 = uniroot(inequality.Q1, interval = search.range, resid = resid, n = n, SET.Alpha1=set.Alpha/2,
      tol=0.001)$root
  Solved.Q2 = NA
  Solved.Q2 = uniroot(inequality.Q2, interval = search.range, resid = resid, n = n, SET.Alpha2=set.Alpha/2,
      tol=0.001)$root
  if (is.na(Solved.Q1)==TRUE) {warning("Unable to solve Q1"); break}
  if (is.na(Solved.Q2)==TRUE) {warning("Unable to solve Q2"); break}
  return(list(Solved.Q1, Solved.Q2))
}

#Defining distributional shift in p as percentage of IC expected value of p
integrand.Ep = function (g0, g1, x, muX, varX) {(1/(sqrt(varX*2*pi)))*exp(g0 + g1*x -0.5*(((x-muX)^2))/varX
    )/(1+exp(g0 + g1*x))}
```

```
integral.Ep = function (g0, g1, muX, varX) {integrate(function (x) {integrand.Ep(g0, g1, x, muX, varX)} , -
    Inf, Inf)$value}
ineq.Ep = function (g0, g1, set.p, muX, varX) {integral.Ep(g0,g1, muX, varX) - set.p}
solve.g0 = function (set.p, g1, muX, varX) {uniroot(ineq.Ep, interval=c(-5,5), g1=g1, set.p=set.p, muX=muX,
     varX=varX)$root}

#Defining distributional shift in lambda as percentage of IC expected value of lambda
integrand.Elamb = function (b0, b1, x, muX, varX) {(1/(sqrt(varX*2*pi)))*exp(b0 + b1*x -0.5*(((x-muX)^2))/
    varX)}
integral.Elamb = function (b0, b1, muX, varX) {integrate(function (x) {integrand.Elamb(b0, b1, x, muX, varX
    )} , -Inf, Inf)$value}
ineq.Elamb = function (b0, b1, set.lamb, muX, varX) {integral.Elamb(b0,b1, muX, varX) - set.lamb}
solve.b0 = function (set.lamb, b1, muX, varX) {uniroot(ineq.Elamb, interval=c(-5,5), b1=b1, set.lamb=set.
    lamb, muX=muX, varX=varX)$root}

#Shewharts ARL1 simulation
Shewhart.OC.RLs <- function(r1, r.mu, r.std, L) {
  n = ncol(r1)
  N = nrow(r1)
  UCL <- r.mu + L*r.std
  LCL <- r.mu - L*r.std
  RLs <- apply(r1, 1, RL, u=UCL, l=LCL, n=n)
  #if (sum(RLs == n) > 0.001*N) {warning("Simulation size too small, in terms of n = nr. of columns") ;
      print(sum(RLs == n))}
  return(RLs)
}

#Procedure to generate ZIP Shewhart ARL1 results for input OC parameters
ARL1.ZI.Shewhart = function (fixed.model, rp.chart, rd.chart, rq.chart, n, N, dist, ICparams , b0.OC = c(),
     g0.OC = c()) {

  #Obtaining chart elements
  L1 = rp.chart[[1]]; mu.rp = rp.chart[[2]]; sd.rp = rp.chart[[3]]
  L2 = rd.chart[[1]]; mu.rd = rd.chart[[2]]; sd.rd = rd.chart[[3]]
  L3 = rq.chart[[1]]; mu.rq = rq.chart[[2]]; sd.rq = rq.chart[[3]]

  #Obtaining IC parameters
  b0 = fixed.model[1]
  b1 = fixed.model[2]
  g0 = fixed.model[3]
  g1 = fixed.model[4]
  muX = fixed.model[5]
  varX = fixed.model[6]
  if (dist == "ZINB") {tau = fixed.model[7]}

  #Define simulation size and output storage
  nr = max((b0.OC), length(g0.OC))
  ARL1.b0.rp = rep(NA, nr)
  ARL1.b0.rd = rep(NA, nr)
  ARL1.b0.rq = rep(NA, nr)
  SDRL1.b0.rp = rep(NA, nr)
  SDRL1.b0.rd = rep(NA, nr)
  SDRL1.b0.rq = rep(NA, nr)
  ARL1.g0.rp = rep(NA, nr)
  ARL1.g0.rd = rep(NA, nr)
  ARL1.g0.rq = rep(NA, nr)
  SDRL1.g0.rp = rep(NA, nr)
  SDRL1.g0.rd = rep(NA, nr)
  SDRL1.g0.rq = rep(NA, nr)

  #Procedure for obtaining OC results
```

```
  idx = 0
  for (k in b0.OC) {
    idx = idx + 1;
    print(eval(sprintf("working on ARL1 simulation %s out of %s", idx, length(b0.OC)+length(g0.OC))))
    if (dist == "ZIP") {Data = Simulate.ZIP.Data(b0 = k, b1 = b1, g0 = g0, g1 = g1, n, N, muX, varX)}
    if (dist == "ZINB") {Data = Simulate.ZINB.Data(b0 = k, b1 = b1, g0 = g0, g1 = g1, tau, n, N, muX, varX)
        }
    Y1 = Data[[1]]
    X1 = Data[[2]]
    rp1 = t(apply(cbind(Y1,X1), 1, pearson.residuals, fixed.model = fixed.model, dist = dist))
    rd1 = t(apply(cbind(Y1,X1), 1, deviance.residuals, fixed.model = fixed.model, dist = dist))
    rq1 = t(apply(cbind(Y1,X1), 1, RQ.residuals, fixed.model = fixed.model, dist = dist))
    RLs = Shewhart.OC.RLs(rp1, mu.rp, sd.rp, L1)
    ARL1.b0.rp[idx] = mean(RLs)
    SDRL1.b0.rp[idx] = sd(RLs)
    RLs = Shewhart.OC.RLs(rd1, mu.rd, sd.rd, L2)
    ARL1.b0.rd[idx] = mean(RLs)
    SDRL1.b0.rd[idx] = sd(RLs)
    RLs = Shewhart.OC.RLs(rq1, mu.rq, sd.rq, L3)
    ARL1.b0.rq[idx] = mean(RLs)
    SDRL1.b0.rq[idx] = sd(RLs)
    print(c(k, ARL1.b0.rp[idx], SDRL1.b0.rp[idx], ARL1.b0.rd[idx], SDRL1.b0.rd[idx], ARL1.b0.rq[idx], SDRL1
        .b0.rq[idx]))
  }
  for (m in g0.OC) {
    idx = idx + 1;
    print(eval(sprintf("Working on ARL1 simulation %s out of %s", idx, length(b0.OC)+length(g0.OC))))
    if (dist == "ZIP") {Data = Simulate.ZIP.Data(b0 = b0, b1 = b1, g0 = m, g1 = g1, n, N, muX, varX)}
    if (dist == "ZINB") {Data = Simulate.ZINB.Data(b0 = b0, b1 = b1, g0 = m, g1 = g1, tau, n, N, muX, varX)
        }
    Y1 = Data[[1]]
    X1 = Data[[2]]
    rp1 = t(apply(cbind(Y1,X1), 1, pearson.residuals, fixed.model = fixed.model, dist = dist))
    rd1 = t(apply(cbind(Y1,X1), 1, deviance.residuals, fixed.model = fixed.model, dist = dist))
    rq1 = t(apply(cbind(Y1,X1), 1, RQ.residuals, fixed.model = fixed.model, dist = dist))
    RLs = Shewhart.OC.RLs(rp1, mu.rp, sd.rp, L1)
    ARL1.g0.rp[idx] = mean(RLs)
    SDRL1.g0.rp[idx] = sd(RLs)
    RLs = Shewhart.OC.RLs(rd1, mu.rd, sd.rd, L2)
    ARL1.g0.rd[idx] = mean(RLs)
    SDRL1.g0.rd[idx] = sd(RLs)
    RLs = Shewhart.OC.RLs(rq1, mu.rq, sd.rq, L3)
    ARL1.g0.rq[idx] = mean(RLs)
    SDRL1.g0.rq[idx] = sd(RLs)
    print(c(m, ARL1.g0.rp[idx], SDRL1.g0.rp[idx], ARL1.g0.rd[idx], SDRL1.g0.rd[idx], ARL1.g0.rq[idx], SDRL1
        .g0.rq[idx]))
  }
  if (length(b0.OC)==length(g0.OC)){
    return(data.frame(b0.OC, ARL1.b0.rp, SDRL1.b0.rp, ARL1.b0.rd, SDRL1.b0.rd, ARL1.b0.rq, SDRL1.b0.rq, g0.
        OC, ARL1.g0.rp, SDRL1.g0.rp, ARL1.g0.rd, SDRL1.g0.rd, ARL1.g0.rq, SDRL1.g0.rq)) }
  if (length(b0.OC)==0){ return(data.frame(g0.OC, ARL1.g0.rp, SDRL1.g0.rp, ARL1.g0.rd, SDRL1.g0.rd, ARL1.g0
      .rq, SDRL1.g0.rq)) }
  if (length(g0.OC)==0){ return(data.frame(b0.OC,ARL1.b0.rp, SDRL1.b0.rp, ARL1.b0.rd, SDRL1.b0.rd, ARL1.b0.
      rq, SDRL1.b0.rq)) }
}
```

C.3 Example of execution: baseline performance evaluation

```
library(VGAM)
```

```
library(statip)
library(pscl)

# BASELINE PERFORMANCE EVALUATION OF THE ZIP REGRESSION-BASED SHEWHART CHARTS
# SYMMETRIC CONTROL LIMITS - IC SCENARIO 1

#Fixed parameters
b0 = 0.1
b1 = 1.0
g0 = 0.5
g1 = -1.0
muX = 0
varX = 1.0

#Fixed model
fixed.model.ZIP.SC1 = c(b0, b1, g0, g1, muX, varX)

#Defining simulation parameters
dist = "ZIP"
N = 10000
n = 3000

#Simulating data for control chart construction: size = Nxn
Data = Simulate.ZIP.Data(b0, b1, g0, g1, n, N, muX, varX)
Y = Data[[1]]
X = Data[[2]]

#Obtain IC residuals for constructing control chart limits
rp0 = t(apply(cbind(Y,X), 1, pearson.residuals, fixed.model = fixed.model.ZIP.SC1, dist = "ZIP"))
rd0 = t(apply(cbind(Y,X), 1, deviance.residuals, fixed.model = fixed.model.ZIP.SC1, dist = "ZIP"))
mu.rp = mean(rp0) ; sd.rp = sd(rp0) ; print(mu.rp) ; print(sd.rp)
mu.rd = mean(rd0) ; sd.rd = sd(rd0) ; print(mu.rd) ; print(sd.rd)


#Solving the chart bounds (This is very slow: 1.5h each in 5000)
L1 = Shewhart.Solve.L(rp0, mu.rp, sd.rp, set.ARL = 200)
L2 = Shewhart.Solve.L(rd0, mu.rd, sd.rd, set.ARL = 200)
L3 = 2.81 ; mu.rq = 0 ; sd.rq = 1 ;

#Storing intermediate results
write.csv(data.frame(L1, L2, L3), file="Baseline-Shewhart-Lvalues-SC1.csv", row.names = FALSE, quote =
    FALSE)

#Calculating desired distributional shift
Elamb.IC = integral.Elamb(b0, b1, muX, varX); Elamb.IC
alpha2 = c(seq(1.0,0.3,-0.1),seq(1.0,3.3,0.3))
dist.shift.lamb = alpha2*Elamb.IC; dist.shift.lamb
b0.OC = sapply(dist.shift.lamb, solve.b0, b1 = b1, muX = muX, varX = varX);
b0.OC[1] = b0; b0.OC[9] = b0; b0.OC
Elamb.OC = sapply(b0.OC, integral.Elamb, b1=b1, muX=muX, varX= varX); Elamb.OC

Ep.IC = integral.Ep(g0, g1, muX, varX) ; Ep.IC
alpha1 = c(seq(1.0,0.3,-0.1), seq(1.0,1.35,0.05)); alpha1
dist.shift.p = alpha1*Ep.IC ; dist.shift.p
g0.OC = sapply(dist.shift.p, solve.g0, g1 = g1, muX = muX, varX = varX)
g0.OC[1] = g0; g0.OC[9] = g0; g0.OC
Ep.OC = sapply(g0.OC, integral.Ep, g1=g1, muX=muX, varX= varX); Ep.OC

#Defining the Shewhart charts
rp.L.Shewhart = list(L1, mu.rp, sd.rp)
rd.L.Shewhart = list(L2, mu.rd, sd.rd)
```

```
rq.L.Shewhart = list(L3, mu.rq, sd.rq)

#Calculating ARL1 results
ARL1.results = ARL1.ZI.Shewhart(fixed.model.ZIP.SC1, rp.L.Shewhart, rd.L.Shewhart, rq.L.Shewhart, n, N,
    dist = "ZIP", ICparams, b0.OC = b0.OC, g0.OC = g0.OC)

#Reordering ARL1 results
ARL1.results$E.lamb.OC = Elamb.OC
ARL1.results$E.p.OC = Ep.OC
ARL1.results = ARL1.results[,c(1,15,2,3,4,5,6,7,8,16,9,10,11,12,13,14)]
k = length(Elamb.OC)
l = length(Ep.OC)
ncol = length(ARL1.results[1,])
ARL1.results = cbind(ARL1.results[1:k,1:(ncol/2)], ARL1.results[(k+1):(k+l),(ncol/2+1):ncol])
ARL1.results

#Storing intermediate results
write.csv(ARL1.results, file="Baseline-Shewhart-L-ARL1-SC1", row.names = FALSE, quote = FALSE)
```

C.4 Functions for GLM-based TBE performance evaluation

```
library(VGAM)
library(statip)
library(pscl)
library(MASS)

#Function to calculate time-between-events
Calculate.TBE = function (YT,r) {
  n = length(YT)
  y = YT[1:(n/2)]
  t = YT[(n/2 + 1):n]
  event.times = t[y>0]
  idx = 1:length(event.times)
  event.times.r = event.times[(idx %% r) == 0]
  tbe = diff(c(0,event.times.r))
  return(tbe)
}

#Function to calculate the accumulated weight, i.e. x, between-events
Calculate.XBE = function (YX,r) {
  n = length(YX)
  y = YX[1:(n/2)]
  x = YX[(n/2 + 1):n]
  idx = 1:(n/2)
  events = idx[y>0]
  Events = rbind(c(0,events[1:(length(events)-1)])+1,events)
  xbe = apply(Events, 2, sumfun, x=x)
  return(xbe)
}

#Function that calculates the number of observations untill the next event
Calculate.nr.obs = function (YX,r) {
  n = length(YX)
  y = YX[1:(n/2)]
  x = YX[(n/2 + 1):n]
  idx = 1:(n/2)
  events = idx[y>0]
  Events = rbind(c(0,events[1:(length(events)-1)])+1,events)
  nr.samples = Events[2,] - Events[1,] + 1
```

```
  return(nr.samples)
}

#Intermediate function to accumulate x values
sumfun = function(x, idx) {
  start = idx[1]
  end = idx[2]
  return(sum(x[start:end]))
}

#Calculating residuals of fitting the IC models on OC data
TBE.raw.residuals = function(model, data) {
  model = TBE.model
  data = cbind(TBE[1,], XBE[1,])
  n = length(data)
  tbe = data[1:(n/2)]
  xbe = data[((n/2) + 1):n]
  mu = predict(model, newdata = data.frame(xbe), type="response")
  raw = TBE - mu
  return(raw)
}

TBE.Pearson.residuals = function(v0, v1, data) {
  n = length(data)
  tbe = data[1:(n/2)]
  xbe = data[((n/2) + 1):n]
  mu = exp(v0 + v1*xbe)
  rp = (tbe - mu) / mu
  return(rp)
}

TBE.Deviance.residuals = function(v0, v1, data) {
  n = length(data)
  tbe = data[1:(n/2)]
  xbe = data[((n/2) + 1):n]
  mu = exp(v0 + v1*xbe)
  t1 =
    d = 2*(-log(tbe/mu) -1 + tbe/mu)
  rd = sign(tbe-mu)*sqrt(d)
  return(rd)
}

TBE.Quantile.residuals = function(v0, v1, shape, data) {
  n = length(data)
  tbe = data[1:(n/2)]
  xbe = data[((n/2) + 1):n]
  mu = exp(v0 + v1*xbe)
  scale = as.numeric(mu/shape)
  cdf = pgamma(tbe, shape=shape, scale=scale)
  rq = qnorm(cdf)
  return(rq)
}

#LI function for individual runs
LI = function(data, q1, q2) {
  len = length(data)/2
  resid = data[1:len]
  accumulated.obs.per.event = data[(len+1):(2*len)]
  n = length(na.omit(resid))
  idx = min(min(n, which(resid > q1)), min(n, which(resid < q2)))
  LI = accumulated.obs.per.event[idx]
```

```
  #if(idx==n){print("Simulation size n too small")}
  return(LI)
}

#Function that obains all LIs for multiple runs
Obtain.LIs = function(resid, alpha, Y, X, r) {
  Q = Solve.Q(resid, alpha)
  Q1 = Q[[1]]
  Q2 = Q[[2]]
  #Calculating nr of observations per event
  nr.obs.per.event = apply(cbind(Y,X), 1, Calculate.nr.obs, r=r)
  accumulated.obs.per.event = lapply(nr.obs.per.event, Reduce, f=sum, accumulate = TRUE)
  accumulated.obs.per.event = do.call(rbind, lapply(accumulated.obs.per.event, function(x) `length<-`(
      unlist(x), ml)))
  #Calculating the LIs
  LIs = apply(cbind(resid, accumulated.obs.per.event), 1, LI, q1 = Q1, q2 = Q2)
  return(LIs)
}

#ALI inequaltiy to solve for alpha
ALI.inequality = function(resid, alpha, Y, X, r, Set.ALI) {mean(Obtain.LIs(resid, alpha, Y, X, r)) - Set.
    ALI}

#Function that solves the required alpha value to assure ALI = Set.ALI (e.g. ALI = 200)
Solve.alpha = function(resid, Y, X, r, Set.ALI) {
  search.range.alpha = c(0, 1)
  alpha = uniroot(ALI.inequality, interval = search.range.alpha, resid = resid, Y=Y, X=X, r=r, Set.ALI=Set.
      ALI, tol=0.001)$root
  return(alpha)
}

#Function to calculate average number of proportion r > q1
Calculate.Alpha1 = function(Y, Q1){
  above.Q1 = apply(Y, 1, function(Y,Q1){sum(Y > Q1, na.rm = TRUE)}, Q1=Q1)
  n = apply(Y, 1, function(x){length(na.omit(x))})
  proportion.above = above.Q1/n
  alpha1 = mean(proportion.above)
  return(alpha1)
}

#Function to calculate average number of proportion r < q2
Calculate.Alpha2 = function(Y, Q2){
  below.Q2 = apply(Y, 1, function(Y,Q2){sum(Y < Q2, na.rm = TRUE)}, Q2=Q2)
  n = apply(Y, 1, function(x){length(na.omit(x))})
  proportion.below = below.Q2/n
  alpha2 = mean(proportion.below)
  return(alpha2)
}

#Inequalities to solve probability control limits, for a given quantile level
inequality.Q1 = function(Y, Q1, n, SET.Alpha1) {Calculate.Alpha1(Y = Y, Q1 = Q1) - SET.Alpha1}
inequality.Q2 = function(Y, Q2, n, SET.Alpha2) {Calculate.Alpha2(Y = Y, Q2 = Q2) - SET.Alpha2}

#Solve the quantile limits Q1 and Q2
Solve.Q = function(Y, set.Alpha){
  #Determining simulation size
  N = nrow(Y)
  #Solving Q1 and Q2
  search.range1 = c(min(Y[Y!=-Inf], na.rm=TRUE), max(Y[Y!=Inf], na.rm=TRUE))
  search.range2 = c(min(Y[Y!=-Inf], na.rm=TRUE), max(Y[Y!=Inf], na.rm=TRUE))
  Solved.Q1 = NA
```

```
  Solved.Q1 = uniroot(inequality.Q1, interval = search.range1, Y = Y, SET.Alpha1=set.Alpha/2, tol=0.001)$
      root
  Solved.Q2 = NA
  Solved.Q2 = uniroot(inequality.Q2, interval = search.range2, Y = Y, SET.Alpha2=set.Alpha/2, tol=0.001)$
      root
  if (is.na(Solved.Q1)==TRUE) {warning("Unable to solve Q1"); break}
  if (is.na(Solved.Q2)==TRUE) {warning("Unable to solve Q2"); break}
  return(list(Solved.Q1, Solved.Q2))
}

#Procedure to generate ZIP Shewhart ALI1 results for input OC parameters
Obtain.LIs.OC.ShewhartQ = function (v0, v1, shape, r, Q1, Q2, Q3, Q4, Q5, Q6, data.dist, n, N, ICparams ,
     b0.OC = c(), g0.OC = c()) {
  #Obtaining IC parameters
  b0 = ICparams[1]
  b1 = ICparams[2]
  g0 = ICparams[3]
  g1 = ICparams[4]
  muX = ICparams[5]
  varX = ICparams[6]
  if (data.dist == "ZINB") {tau = ICparams[7]}

  #Define simulation size and output storage
  n = ncol(rp0)
  N = nrow(rp0)
  nr = max((b0.OC), length(g0.OC))
  Store.LI = matrix(rep(NA, N*(length(b0.OC)+length(g0.OC))*3),N,(length(b0.OC)+length(g0.OC))*3)

  #Procedure for obtaining OC results
  idx = 0
  for (k in b0.OC) {
    idx = idx + 1;
    print(eval(sprintf("working on ALI1 simulation %s out of %s", idx, length(b0.OC)+length(g0.OC))))
    if (data.dist == "ZIP") {Data = Simulate.ZIP.Data(b0 = k, b1 = b1, g0 = g0, g1 = g1, n, N, muX, varX)}
    if (data.dist == "ZINB") {Data = Simulate.ZINB.Data(b0 = k, b1 = b1, g0 = g0, g1 = g1, tau, n, N, muX,
        varX)}
    Y1 = Data[[1]]
    X1 = Data[[2]]
    arrival.times = t(matrix(rep(1:n),n,N))

    #Computing the time and weights between events
    TBE1 = apply(cbind(Y1,arrival.times), 1, Calculate.TBE, r=r)
    XBE1 = apply(cbind(Y1,X1), 1, Calculate.XBE, r=r)
    ml1 = max(lengths(TBE1))
    TBE1 = do.call(rbind, lapply(TBE1, function(x) `length<-`(unlist(x), ml1)))
    XBE1 = do.call(rbind, lapply(XBE1, function(x) `length<-`(unlist(x), ml1)))

    #obtaining residuals
    rp1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Pearson.residuals, v0 = v0, v1 = v1))
    rd1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Deviance.residuals, v0 = v0, v1 = v1))
    rq1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Quantile.residuals, v0 = v0, v1 = v1, shape = shape))

    #Calculating the LIs
    nr.obs.per.event = apply(cbind(Y1,X1), 1, Calculate.nr.obs, r=r)
    accumulated.obs.per.event = lapply(nr.obs.per.event, Reduce, f=sum, accumulate = TRUE)
    accumulated.obs.per.event = do.call(rbind, lapply(accumulated.obs.per.event, function(x) `length<-`(
        unlist(x), ml1)))

    LIs.rp.b = apply(cbind(rp1, accumulated.obs.per.event), 1, LI, q1 = Q1, q2 = Q2)
    LIs.rd.b = apply(cbind(rd1, accumulated.obs.per.event), 1, LI, q1 = Q3, q2 = Q4)
    LIs.rq.b = apply(cbind(rq1, accumulated.obs.per.event), 1, LI, q1 = Q5, q2 = Q6)
```

```
    Store.LI[1:N,(3*(idx-1)+1)] = LIs.rp.b
    Store.LI[1:N,(3*(idx-1)+2)] = LIs.rd.b
    Store.LI[1:N,(3*(idx-1)+3)] = LIs.rq.b
  }
  for (m in g0.OC) {
    idx = idx + 1;
    print(eval(sprintf("Working on ALI1 simulation %s out of %s", idx, length(b0.OC)+length(g0.OC))))
    if (data.dist == "ZIP") {Data = Simulate.ZIP.Data(b0 = b0, b1 = b1, g0 = m, g1 = g1, n, N, muX, varX)}
    if (data.dist == "ZINB") {Data = Simulate.ZINB.Data(b0 = b0, b1 = b1, g0 = m, g1 = g1, tau, n, N, muX,
        varX)}
    Y1 = Data[[1]]
    X1 = Data[[2]]
    arrival.times = t(matrix(rep(1:n),n,N))

    #Computing the time and weights between events
    TBE1 = apply(cbind(Y1,arrival.times), 1, Calculate.TBE, r=r)
    XBE1 = apply(cbind(Y1,X1), 1, Calculate.XBE, r=r)
    ml1 = max(lengths(TBE1))
    TBE1 = do.call(rbind, lapply(TBE1, function(x) `length<-`(unlist(x), ml1)))
    XBE1 = do.call(rbind, lapply(XBE1, function(x) `length<-`(unlist(x), ml1)))

    #obtaining residuals
    rp1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Pearson.residuals, v0 = v0, v1 = v1))
    rd1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Deviance.residuals, v0 = v0, v1 = v1))
    rq1 = t(apply(cbind(TBE1,XBE1), 1, TBE.Quantile.residuals, v0 = v0, v1 = v1, shape = shape))

    #Calculating the LIs
    nr.obs.per.event = apply(cbind(Y1,X1), 1, Calculate.nr.obs, r=r)
    accumulated.obs.per.event = lapply(nr.obs.per.event, Reduce, f=sum, accumulate = TRUE)
    accumulated.obs.per.event = do.call(rbind, lapply(accumulated.obs.per.event, function(x) `length<-`(
        unlist(x), ml1)))

    LIs.rp.g = apply(cbind(rp1, accumulated.obs.per.event), 1, LI, q1 = Q1, q2 = Q2)
    LIs.rd.g = apply(cbind(rd1, accumulated.obs.per.event), 1, LI, q1 = Q3, q2 = Q4)
    LIs.rq.g = apply(cbind(rq1, accumulated.obs.per.event), 1, LI, q1 = Q5, q2 = Q6)
    Store.LI[1:N,(3*(idx-1)+1)] = LIs.rp.g
    Store.LI[1:N,(3*(idx-1)+2)] = LIs.rd.g
    Store.LI[1:N,(3*(idx-1)+3)] = LIs.rq.g
  }
  return(data.frame(Store.LI))
}
```

C.5 Example of execution: TBE performance evaluation

```
library(VGAM)
library(statip)
library(pscl)
library(MASS)

#True parameters
b0 = 0.1
b1 = 1.0
g0 = 0.5
g1 = -1.0
muX = 0
varX = 1

#Size of the phase I data
m = 1500
```

```
#Forloop to eliminate the variation that originates from having a phase I
nr.loops = 100
Qvalues = matrix(NA, nr.loops,6)
for (i in 1:nr.loops) {
print(i)

#Simulating phase 1 data - Y = response variable, X = covariate - size = 1 x m
Data = Simulate.ZIP.Data(b0, b1, g0, g1, m, 1, muX, varX)
Y = as.numeric(Data[[1]])
X = as.numeric(Data[[2]])
arrival.times = rep(1:m)
r = 1

#Calculating the TBE and XBE from the phase 1 data
tbe = Calculate.TBE(cbind(Y, arrival.times), r=r)
xbe = Calculate.XBE(cbind(Y,X), r=r)

#Obtaining the phase 1 TBE model
TBE.model = glm(tbe ~ xbe, family=Gamma(link="log"))
coef = coefficients(TBE.model)
v0 = coef[1]
v1 = coef[2]
shape.estimation = gamma.shape(TBE.model)
est.shape = as.numeric(shape.estimation[1])

#Defining simulation parameters
dist = "ZIP"
N = 200
n = 3000

#Simulating data for control chart construction: size = Nxn
Data = Simulate.ZIP.Data(b0, b1, g0, g1, n, N, muX, varX)
Y = Data[[1]]
X = Data[[2]]
arrival.times = t(matrix(rep(1:n),n,N))

#Computing the time and weights between events
TBE = apply(cbind(Y,arrival.times), 1, Calculate.TBE, r=r)
XBE = apply(cbind(Y,X), 1, Calculate.XBE, r=r)

#Converting into proper format
ml = max(lengths(TBE))
TBE = do.call(rbind, lapply(TBE, function(x) `length<-`(unlist(x), ml)))
XBE = do.call(rbind, lapply(XBE, function(x) `length<-`(unlist(x), ml)))

#Fitting a ZIP GLM to the data and obtain residuals (15 min)
rp0 = t(apply(cbind(TBE,XBE), 1, TBE.Pearson.residuals, v0 = v0, v1 = v1))
rd0 = t(apply(cbind(TBE,XBE), 1, TBE.Deviance.residuals, v0 = v0, v1 = v1))
rq0 = t(apply(cbind(TBE,XBE), 1, TBE.Quantile.residuals, v0 = v0, v1 = v1, shape = est.shape))
print(sum(rq0==Inf,na.rm=TRUE)); rq0[rq0==Inf] = max(rq0[rq0!=Inf],na.rm=TRUE)

#Solving the control limits such that ALI = 200
alpha.rp0 = Solve.alpha(rp0, Y, X, r, Set.ALI = 200); print(alpha.rp0)
alpha.rd0 = Solve.alpha(rd0, Y, X, r, Set.ALI = 200); print(alpha.rd0)
alpha.rq0 = Solve.alpha(rq0, Y, X, r, Set.ALI = 200); print(alpha.rq0)
Q = Solve.Q(rp0, alpha.rp0); Q1 = Q[[1]]; Q2 = Q[[2]]
Q = Solve.Q(rd0, alpha.rd0); Q3 = Q[[1]]; Q4 = Q[[2]]
Q = Solve.Q(rq0, alpha.rq0); Q5 = Q[[1]]; Q6 = Q[[2]]

#Storing intermediate results
```

```
Qvalues[i,] = c(Q1, Q2, Q3, Q4, Q5, Q6)

#Calculating desired distributional shift
Elamb.IC = integral.Elamb(b0, b1, muX, varX); Elamb.IC
alpha2 = c(seq(1.0,0.3,-0.1),seq(1.0,3.3,0.3))
dist.shift.lamb = alpha2*Elamb.IC; dist.shift.lamb
b0.OC = sapply(dist.shift.lamb, solve.b0, b1 = b1, muX = muX, varX = varX);
b0.OC[1] = b0; b0.OC[9] = b0; b0.OC
Elamb.OC = sapply(b0.OC, integral.Elamb, b1=b1, muX=muX, varX= varX); Elamb.OC

Ep.IC = integral.Ep(g0, g1, muX, varX) ; Ep.IC
alpha1 = c(seq(1.0,0.3,-0.1), seq(1.0,1.35,0.05)); alpha1
dist.shift.p = alpha1*Ep.IC ; dist.shift.p
g0.OC = sapply(dist.shift.p, solve.g0, g1 = g1, muX = muX, varX = varX)
g0.OC[1] = g0; g0.OC[9] = g0; g0.OC
Ep.OC = sapply(g0.OC, integral.Ep, g1=g1, muX=muX, varX= varX); Ep.OC

#Calculating ARL1 results
ICparams = c(b0, b1, g0, g1, muX, varX)
LIs = Obtain.LIs.OC.ShewhartQ(v0 = v0, v1 = v1, shape = est.shape, r, Q1, Q2, Q3, Q4, Q5, Q6, data.dist="
     ZIP", n, N, ICparams, b0.OC = b0.OC, g0.OC = g0.OC)

#Merge results
if (i == 1) {current.LIs = LIs}
if (i > 1) {current.LIs = rbind(current.LIs,LIs)}
}

write.csv(current.LIs, file="All-LIs-TBE-SC1.csv", row.names = FALSE, quote = FALSE)
ALI1.results = merge.results(LIs, pooled.sd = FALSE)

#Reordering ALI1 results
ALI1.results$b0.OC = b0.OC
ALI1.results$g0.OC = g0.OC
ALI1.results$E.lamb.OC = Elamb.OC
ALI1.results$E.p.OC = Ep.OC
ALI1.results = ALI1.results[,c(13,15,1,2,3,4,5,6,14,16,7,8,9,10,11,12)]
print(ALI1.results)

#Storing results
write.csv(ALI1.results, file="TBE-ALI1-SC1", row.names = FALSE, quote = FALSE)
write.csv(Qvalues, file="TBE-Qvalues-SC1.csv", row.names = FALSE, quote = FALSE)
```