Statistical and Predictive Process Monitoring

How to Monitor Complex Processes in the Age of Big Data?

Leo Huberts, section Operations Management (ABS/UvA)

Defense date: April 22, 2021

Promotor: Prof. dr. R.J.M.M. Does (UvA)
Co-promotor: Dr. M. Schoonhoven (UvA)

In this dissertation, Statistical Process Monitoring (SPM) and Predictive Process Monitoring (PPM) for big data sets have been discussed. We analyzed the use of classical control charting techniques as well as predictive solutions.

In Chapter 2 of this thesis, we investigated the use of the Central Limit Theorem (CLT) in monitoring large data streams. Because averages are normally distributed under certain conditions, according to the CLT, this should largely resolve the issue of non-normally distributed data. However, we showed that the tail behavior for the means of non-normally distributed subsamples deviates strongly from normality. The degree to which the distribution of the mean deviates depends on various factors: the sample size, the number of samples, the specified desired performance of the control chart, and the degree of the deviation from normality. For example, when the deviation from normality is substantial due to heavy tails (t_4) or substantial skewness (lognormal), the tail behavior cannot be accurately approximated by the normal distribution even when the sample size is 1000. The implications are especially relevant for process monitoring.

Chapter 3 of this thesis is concerned with the continuous updating of parameters during process monitoring. We studied the effects of updating in various scenarios for three types of control charts. The results support updating control limits as long as the reason for out-of-control signals is known and the origin can be retraced. If this is not the case, the best strategy depends on the size of the expected mean deviation. We suggest further research on the behavior of updating the limits for various subgroup sample sizes, as well as on performance for varying distributional assumptions.

In Chapter 4, a procedure to introduce a delay in updating control chart parameters is discussed. As discussed in Chapter 3, updating using contaminated samples should be avoided. The methods described in this chapter prevent these contaminated updates while maintaining the improvements in parameter estimation. In a case study using COVID-19 related data, we demonstrated the added value of updating control chart parameters for mortality rates in the Netherlands.

The second part of this thesis considers PPM. In Chapter 5 we considered the use of various machine learning techniques in PPM. A wide range of predictive techniques is available that are largely data-driven. We introduce a procedure to tune these predictions towards a desired false alarm rate in monitoring. Using a unique non-public data set on mental health, we investigate the performance of machine learning techniques. The Extreme Gradient Boosting (XGBoost) algorithm is subsequently used to monitor the risk of relapse in people diagnosed

with schizophrenia. The procedure can aid healthcare workers in identifying people that are likely to need preventive care. Future research using more consistent data and a longer timeframe is encouraged. Neural networks can potentially improve predictions, as well as the addition of high-frequency data sources.

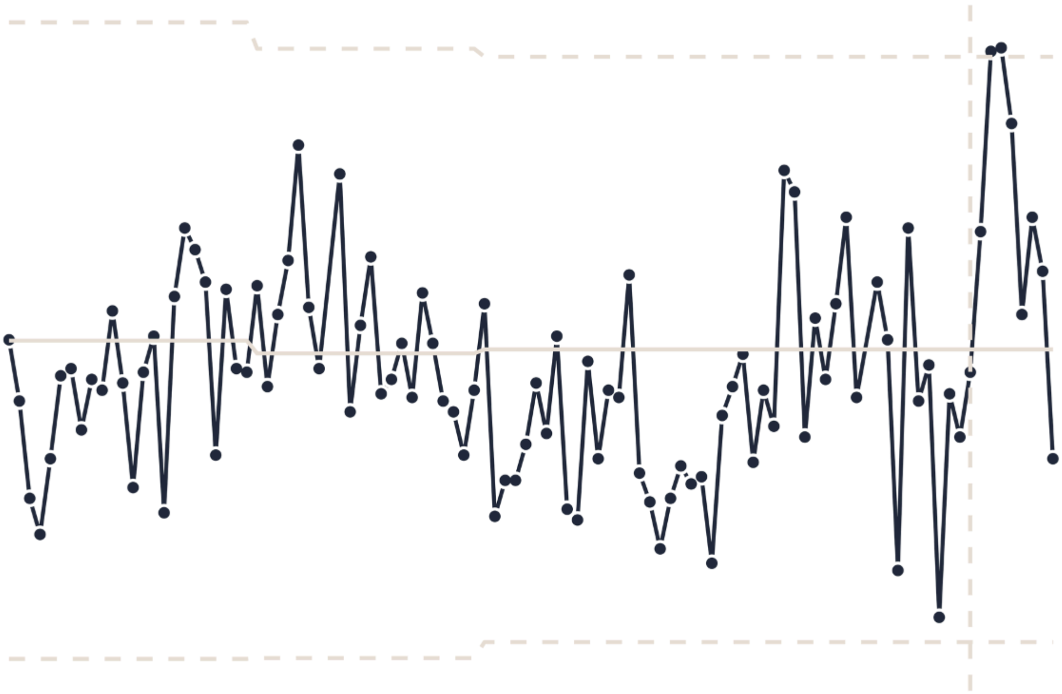
In Chapter 6 of this thesis, we introduced multilevel process monitoring. Modeling the hierarchical structure of a process can improve parameter estimates and the predicted probabilities. Furthermore, using a multilevel model allows monitoring at the different measurement levels. An educational case study was presented to illustrate this approach. Bayesian hierarchical modeling was used in a predictive monitoring procedure. This method produced more accurate predictions than the appropriate machine learning method. The procedure allows early warnings for students that have 'exceptional' performance. This aids schools in personalizing education and quality control. We suggest further research of the procedure using industrial process data of a hierarchical nature and varying the Bayesian priors in analyses.

In conclusion, the increase in available data and improvements in technology enable a new phase in SPM and PPM. Updating process parameter estimates will improve the use of control charts. Introducing a delay in these updates can prevent the use of contaminated data. Furthermore, early intervention based on PPM in services and industry can support the efficient use of resources and prevent processes and people from spiraling out of control.

STATISTICAL AND PREDICTIVE

PROCESS MONITORING

LEO C.E. HUBERTS



MONITORING COMPLEX PROCESSES IN THE AGE OF BIG DATA