Chapter 1

Data

In this chapter, we will introduce the pharmaceutical production data, that we shall later use to infer a causal structure pertaining different parts of the production system. In particular, as we are interested in the duration and amount of produced substance during the production flow, these are highly relevant attributes of the processes that make up the production. Hence, we will start this chapter with an overview of the production system, how the observations are structured and created to begin with. For the rest of the chapter, we will concern ourselves with analysis of the production system such as basic statistics, incomplete or wrongly labelled observations and initial observations about codependency (which is very relevant when studying causality).

The observations that we will ultimately use for the causal study is simulated by [?]. However, before diving into how these simulations were carried out, we present the overall structure of the simulated observations and the production system they are supposed to originate from. Namely, a set of 6 cycles, where each cycle consists of a set of batches executed one by one. Thus, as cycle is simply a notion for multiple batches that are executed in continuation of each other. In particular, different settings for the simulation of each cycle have been used to encompass multiple scenarios of how the production system can function. We note that although the cycles are generated from different settings, they are still representative of the same production system. Hence, we shall later combine observations from all cycles.

A batch refers to a collection of processes/unit $\mathcal{U}$ that need to be executed in some order to produce a product. In particular, for this simulation study, each batch is a collection of the processes depicted in Figure 1.1.

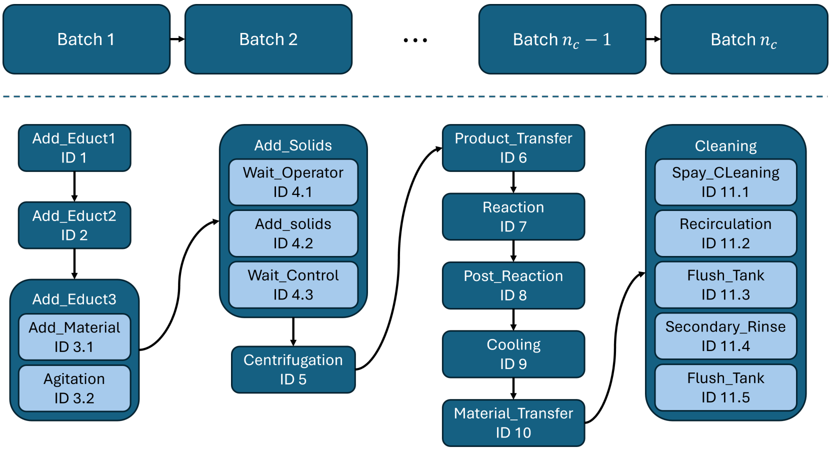

Figure 1.1: The structure of a single cycle. A cycle comprises n_c batches that are carried out one by one. Thus, a cycle is simply a collection of a complex task (a batch), that need to be repeated n_c times. The number of times is often based on time or amount of produced drug or substance, such that a cycle is terminated after these criterions are met. We shall later see that in the case of this simulation study, n_c is determined from the accumulated duration of the batches i.e. after a certain amount of time has been simulated, the simulation is terminated. The structure of each batch is observed in the lower part of the figure and consists of 11 main processes such as addition of solids and chemicals (processes labeled with ID 1 through 4). Each process can be made up of a number of *subprocesses* such as the subprocess with ID 4.3, where the batch waits for a control operator before proceeding.

For each cycle, we then have a time series, where the ID of the batch is given as well as sensory values. The sensors measure the level of the tank (percentage of mow much of the tank is filled), the height (equivalent to the level), the RPM of a motor that circulates the contents of the tank, the cooling water flow (specifically for the cooling process with ID 9) and the steam flow during the reaction process. We shall however restrict ourselves to only using the level sensor in this thesis but including the other variables could improve on our results later on. We have chosen only the level as it is assumed to be the most descriptive of how much product is eventually produced.

In Figure 1.2 an example of the temporal evolution of a process is shown (with the previous process as well). We define T_u^P to be the duration of the process/unit operation u and equivalently M_u^P to be the *change* in level during process u. Not that we have also defined random variables T_u^D and M_u^D. Why we need these will become apparent in Section 1.3 but for now we note that they correspond to delays after each of the processes. In particular, after a process is completed, the might be some downtime in the production system due to unforeseen reasons. We shall later see that for some processes, the delays will not influence the level of the tank whereas the reverse is true for other processes such as the reaction (labelled 7).

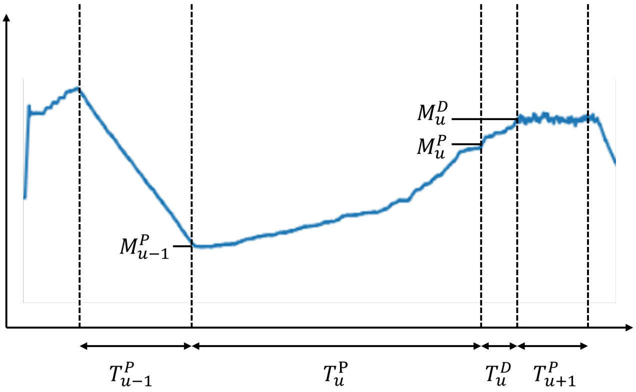

Figure 1.2: Exemplification of the evolution of the level in a tank during a process u and the previous process. The variables T_u^P, T_u^D, M_u^P and M_u^D related to the process are shown. Note that M_u^P and M_u^D are the changes in level from the previous process or delay of process such that they describe the accumulated evolution in the level of the tank during said process. In particular, changes in levels can occur when the production system is idle.

We note that simulations were carried out through a mixture of `Simulink` and `Stateflow` simulations. In particular, the continuous subsystems such as the reaction in process 7 where simulated through `Simulink` based mass-balance equations.

At this point, we have a rudimentary understanding of how the system is simulated and the meaning of the random variables that are related to each process. We thus continue with some basic statistics concerning the durations of batches. For the remaining of the chapter, we will primarily present results for the duration and delays of the processes as the analysis and results are identical to those of the change in level. Namely, we shall observe that the dimension of the random vector that describes each batch (i.e. durations, delays and level changes) large enough such that no meaningful conclusion on the causal relation between random variables can be drawn. In particular, we will need a framework such as the one presented in **??**, to discover such relationships.

1.1 Basic statistics

Before analyzing the time series in more depth and filter out (or correct) troublesome data points, we present some initial statistics on the duration of batches for each cycle. The statistics are summarized in Table 1.1 below. We note that some difference is observed from cycle to cycle, but we choose to assume that these differences are simply a feature of the production system, such that later on, we can combine all observations across all cycles into a single dataset to be used for causal discovery.

Cycle	number of batches	mean	variance	standard deviation	coefficient of variation
A	66	14.776	3.641	1.908	0.1291
B	64	15.644	3.915	1.979	0.1265
C	61	17.714	2.330	1.526	0.08617
D	60	18.069	6.922	2.631	0.1456
E	60	18.088	9.613	3.100	0.1714
F	63	17.227	7.766	2.787	0.1618
Combined cycles	374	16.876	7.218	2.687	0.1592

Table 1.1: Basic batch statistics for each cycle and by combining all cycles into a single data set. The average duration of batches across cycles appear similar when taking the variance of the durations into account. We note that later, we wish to estimate the dependency between pairs of random variables whence more observations is better, as always in data science. We do however note that there appears to be a difference between especially the first three cycles and the latter ones. In particular, the variance is larger for cycles named D, E and F. The source of this variation is at this point unknown however it could be seen as a feature of the dataset. Namely, if the observations are truly from the same production system, this variation could be an inherent feature of the production system which we should not remove.

In the following section, we discuss a problem with some of the batches. Namely, the trailing batches, which appear to be cut-off during simulation. In this way, we shall end up with a total of 368 batches, which after some correction (see Section 1.3) will be our final data set.

1.2 Incompleteness of trailing batches

In this section, we shall investigate the combined dataset of 374 batches in more detail. In particular, we shall observe some deviation from Figure 1.1 in terms of labels of each event in the time series and how we have handled these discrepancies. Namely, by looking through the time series for each cycle, we observe entries labeled with negative processes. These, we will investigate the next section and note that from paper introducing the simulations we present here [?], it is by design that some labels are incorrect. Their argument for this is in relation to training a robust machine learning algorithm but as this is none of our concern, we shall manually handle these in correct labels. In particular, the negative labels are initially negated to be positive instead. Hence, we observe events labeled 3, which is not originally a part of the production system description from Figure 1.1. With these negative labels transformed, we count for each of the (new) process labels, how many batches are observed. E.g. how many batches are at some point observed to be undergoing process 1 (the addition of a material). We do this to make sure that in fact every batch go through all processes from Figure 1.1. The result of this counting batches is presented in Table 1.2, where the description of the recognized processes has been copied from [?]. Note that `Educt1`, `Educt2` and `Educt3` are just some (unknown) materials that we do not care about. Note that we have not included labels pertaining the cleaning operation as these will be handled separately in

ID	Count	Description
1.0	374	Addition of liquid raw material `Educt1`
2.0	374	Addition of liquid raw material `Educt2`
3.0	181	
3.1	374	Addition of liquid raw material `Educt3`
3.2	374	Agitation
4.0	163	
4.1	374	Waiting for field operation
4.2	374	Addition of solids
4.3	374	Waiting for control operator
5.0	374	centrifugation
6.0	374	Product transfer
7.0	370	Reaction
8.0	369	Post reaction
9.0	369	Cooling
10.0	368	Material transfer

Table 1.2: The number of batches across all cycles that contains at least one observation for each different process label.

Section 1.4 where we also argue why we will not use these observations in the later analysis.

Interestingly, the *unrecognized* process labels 3 and 4 only occur for processes with subprocesses. We shall later observe that these labels all originate from negative process labels and that they actually correspond to delays between processes as portrayed in Figure 1.2. For now, we however concentrate on the last four process labels 7, 8, 9 and 10. In particular, as all the other process labels (excluding 3 and 4) appear exactly 374 (the number of batches in total) times, we suspect that something weird is going on with these *missing* observations. As hinted to before, it turns out that the simulations have been cut off after 1100 hours. Therefore, the trailing batch of each cycle does not complete all processes. For example, in Figure 1.3, we have shown the first batch of cycle A as well as the trailing batch and how the over time switch between process labels (not that for this plot, we have not negated the negative process labels)

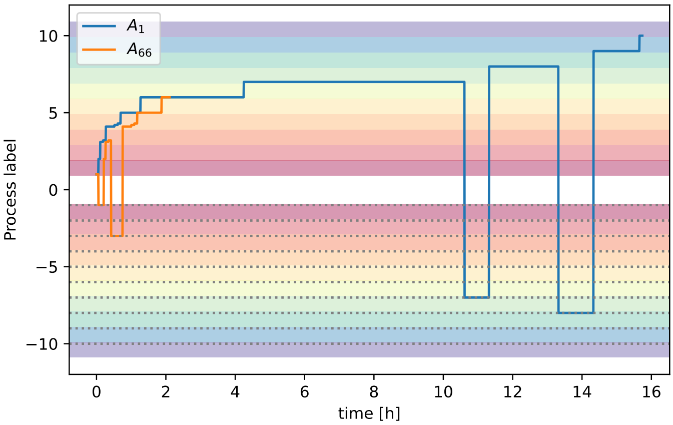

Figure 1.3: The first and last batch from cycle A. The horizontal colored bars corresponds to the different process labels such that time stamps labeled e.g. 3.1 and 3.2 fall in the same colored region. It is clear that the final batch is cut-off before finishing the last process. Furthermore, we observe that the negative process labels for these two batches only occur before the process label enters a new colored region. This hints to the negative process labels are actually delays between processes.

From the above, it is clear that we need to remove the final batches of each cycle. Thus, we now have a total of 368 batches. In the following section, we shall see in more detail when the negative process labels occur and make the assumption that they correspond to delays between processes. Note that the cleaning operation is not considered in the following section.

1.3 Production processes

We now focus on the processes labelled 1 through 10 from Figure 1.1. In particular, we shall denote these processes as *production* processes, as they are exactly the processes where a substance is produced or handled in some other way. Initially, we shall however focus on the first processes up to and including 4.3. Namely, from Table 1.2, we saw that it was these few initial processes where labels seemed to be weird.

In Figure 1.4, we have shown the 22$^{\text{nd}}$ batch of cycle B. Once again, we observe the negative process label. We notice that it is only visited once, and only at the of the process which its label corresponds to.

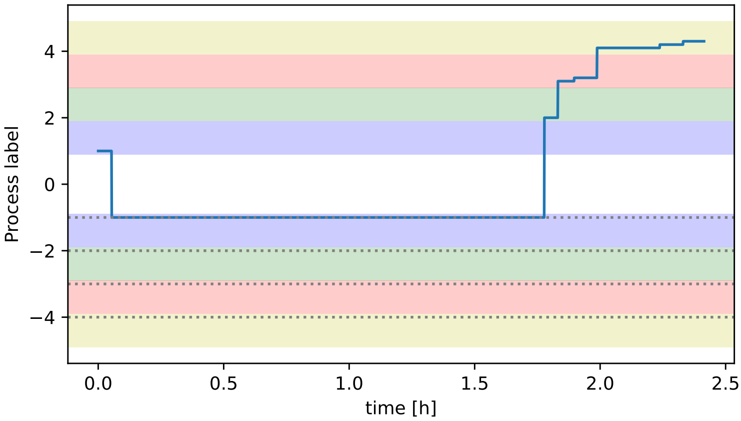

Figure 1.4: The temporal evolution of process labels for batch 22 from cycle B. Only the processes pertaining to the first boxes of Figure 1.1 are shown to easier tell what is happening.

Continuing the investigating, we see that negative process labels occur throughout all the six cycles. Furthermore, by saving what negative process labels have occurred for each cycle, we obtain Table 1.3 where we observe a clear tendency regarding the process labels −4.1, −4.2 and −4.3. Namely, they only occur in cycle F. From [?], we note that cycle F is the only phase containing wrongly labeled time points. In particular, we can conclude that the negative process labels apart from −4.1, −4.2 and −4.3 are not an error in the data set.

In Figure 1.5, we have shown some of the batches which contain the process labels −4.1 etc. We observe that if either of the three process labels are negative, then all of them are and no corresponding positive labels occur. We shall thus

Cycle Event	A	B	C	D	E	F
-1	■	■	■	■		
-2				■	■	■
-3	■		■	■	■	■
-4		■	■	■	■	
-4.1						■
-4.2						■
-4.3						■
-5	■	■	■	■	■	■
-6		■	■	■	■	■
-7	■		■	■	■	■
-8	■		■	■	■	■
-9				■	■	■
-10		■		■	■	■

Table 1.3: Occurrences of negative process labels. It is observed that the process labels -4.1, -4.2, -4.3 only occur in cycle F which is known to be the only cycle with wrongly labelled phases.

assume that whenever -4.1, -4.2 or -4.3 is observed, it is actually just the negated process label. Correcting the data set under this assumption, we then only have negative process labels that are integer which we have summarized in Table 1.4 below.

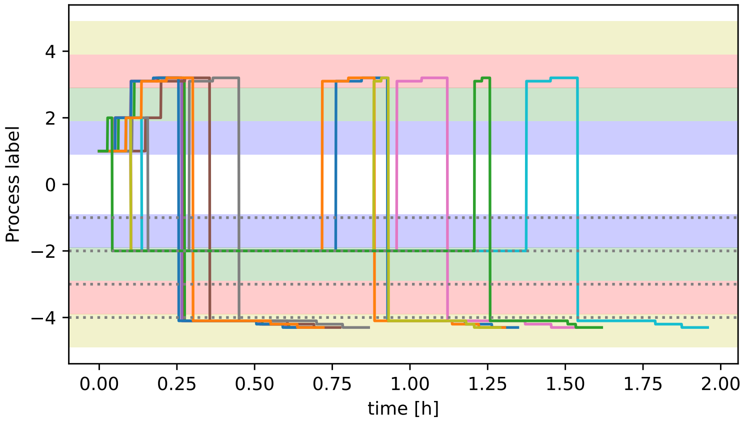

Figure 1.5: 13 out of the total 48 batches where at least one of the process labels -4.1, -4.2 or -4.3 were observed.

Cycle Event	A	B	C	D	E	F
-1	■	■	■	■		
-2				■	■	■
-3	■		■	■	■	■
-4		■	■	■	■	
-5	■	■	■	■	■	■
-6		■	■	■	■	■
-7	■		■	■	■	■
-8	■		■	■	■	■
-9				■	■	■
-10		■		■	■	■

Table 1.4: In the modified data set, where −4.1, −4.2 and −4.3 have been converted their absolute value. We observe that the occurrence of negative labels is not identical from cycle to cycle. Depending on the parameters of the simulation, this could either be per happenstance on different settings of the simulation. Either way, we assume that as the simulation is based on the same production system, this variation is observed naturally. Hence, we shall not do more with these observations in terms of filtering them out or correcting them.

From Table 1.4, we see that cycles D, E and F appear to contain more negative process labels. We will however assume that the cycles are simulated from the same production system such that no hyperparameters were changes. In particular, we shall assume that the observations of negative process labels occurs at random, independently of the cycle.

Furthermore, by plotting the different batches from different cycles, it is apparent that the negative process labels always occur after each process (including subprocesses) and before the next process. I.e. we only observe the label −1 after 1 and before 2. Likewise, −3 only occurs after both 3.1 and 3.2 but before 4.1. As hinted to before, we shall thus assume that these negative process labels corresponds to delays between processes. This does make sense from a production point, but they also note in [?] that delays between operations have been implemented.

At this point, it would seem that the labels of the processes are understood for phases 1 through 10 corresponding to the actual production in each batch. Thus, we proceed by searching relationships and otherwise quantifying the durations of each phase, both delays and duration for each of the phases. As a beginning, histograms for each of the phases and delays are plotted in Figure 1.6. Notice that phases 4.1, 4.3 and 8 are not shown, this is because they always last 15 min, 5 min and 2 hours respectively with the only derivation being in machine precision either when loaded or during calculations. Furthermore, notice that for the negative IDs i.e. the delays, the orange bar. This bar represents the cases where no delay was observed which is thus modelled as an atom at 1.

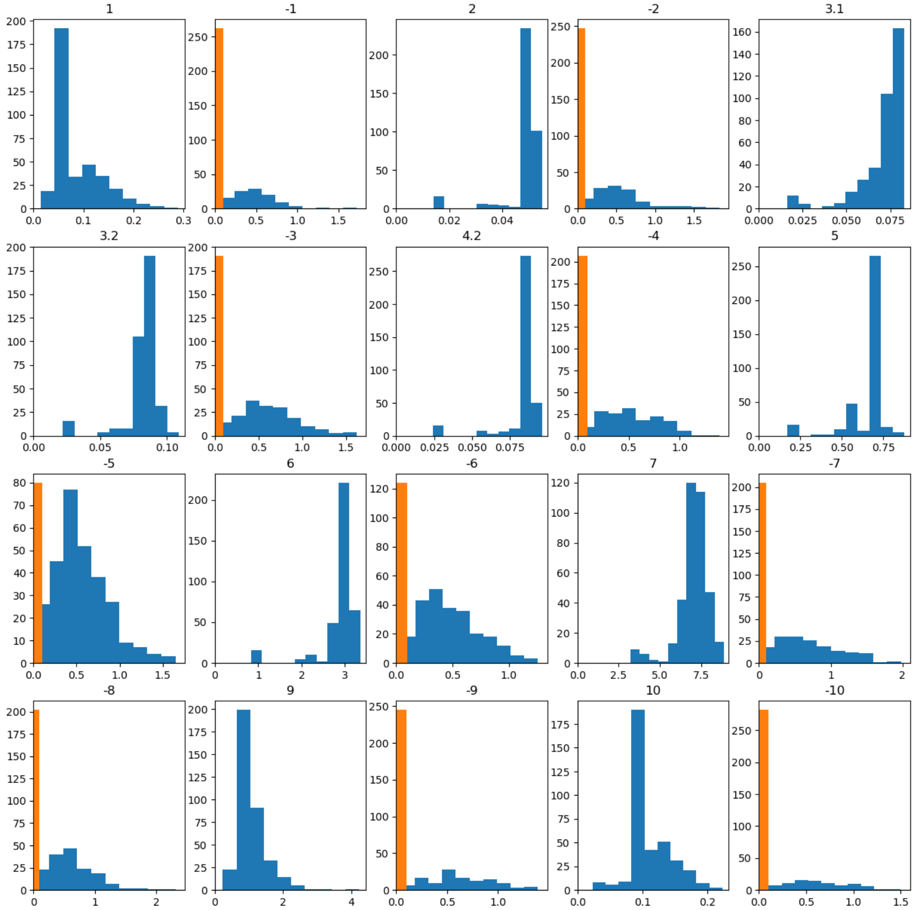

Figure 1.6: Histograms of all phases and delays which are non-constant.

Apart from the above comments, not much catches the eye when looking at Figure 1.6, and we thus proceed by checking if any correlation is immediately present.

Cycle	A	B	C	D	E	F
$\widehat{\mathbb{E}\left[\sum X_u\right]}$	13.993	13.898	15.343	14.471	14.589	14.418
$\widehat{\mathrm{Var}\left(\sum X_u\right)}$	0.95636	0.46587	0.76111	4.9589	4.2678	5.3545
$\sum \widehat{\mathrm{Var}\left(X_u\right)}$	0.50590	0.31182	0.36667	1.8322	1.5788	1.9696
$\widehat{\mathbb{E}\left[\sum X_u^D\right]}$	0.96398	1.9402	2.4503	3.6050	3.7390	3.0041
$\widehat{\mathrm{Var}\left(\sum X_u^D\right)}$	0.31843	0.39117	0.90187	1.2468	1.2787	1.0462
$\sum \widehat{\mathrm{Var}\left(X_u^D\right)}$	0.34921	0.53198	0.74914	1.4357	1.2454	1.3099
$\widehat{\mathbb{E}\left[\sum X_u X_u^D\right]}$	1.9321	1.5001	4.8191	6.4225	6.0405	6.3343
$\widehat{\mathrm{Var}\left(\sum X_u X_u^D\right)}$	3.7798	0.89920	16.870	22.133	12.660	16.194

Table 1.5: Each of the time related variables X_i and D_i and variance description.

1.3.1 Correlations

Lige en korrelationsmatrix

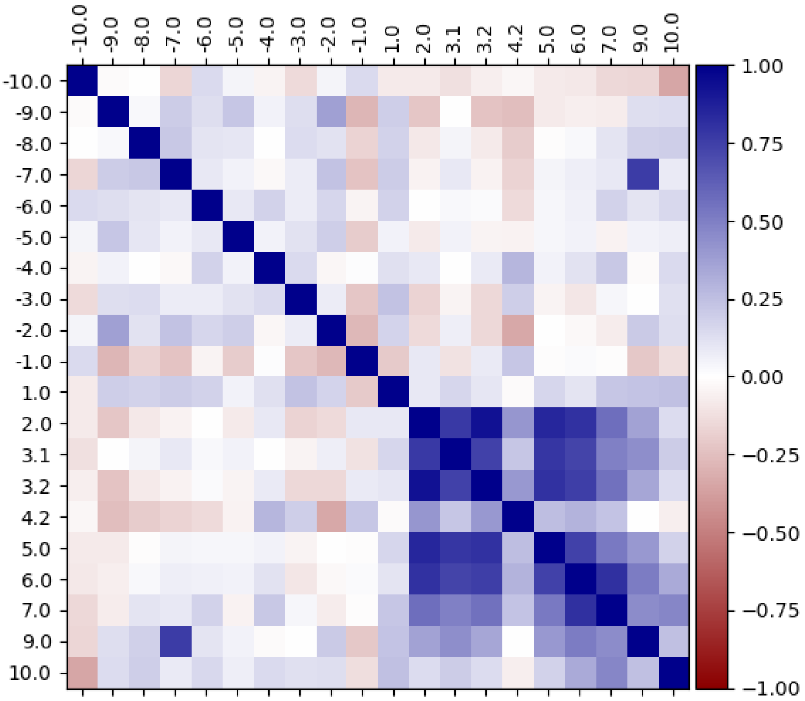

Figure 1.7: Correlation matrix for all phases with non-constant duration.

9 og 10 er ikke specielt korreleret med noget (afkøling og materiale overførsel). Ellers er 2 fremt il og med reaktionen alle korrelerede med hinanden. Eftersom rent fysisk det udvikles i tid, må handlingen i 2 påvirke de næste osv.

Umiddelbart lidt spøjst hvis delay på 7 (reaktion) skulle have noget med tiden for afkøling at gøre, især at den skulle være positiv (ville man ikke tro delay efter produktion ville afkøle mere og dermed reducere behov for afkøling, medmindre varmt steam bliver tilføjet også under delay på 7)

Herunder er samme korrelationsmatrix, dog hvor delay og phasens varighed lagt sammen (også med sub phases såsom 3.1, 3.2 og -3 tilsammen bliver 3)

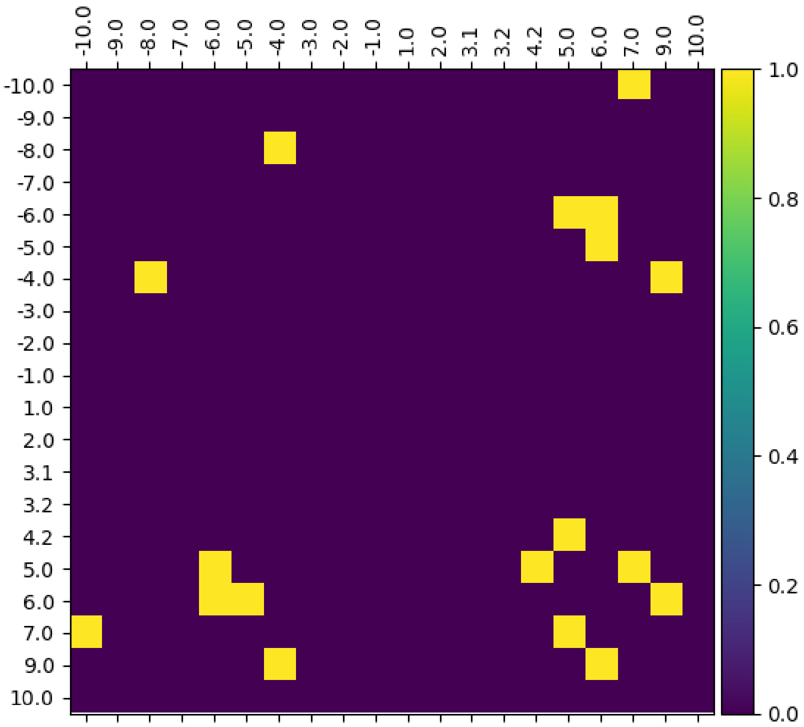

Figure 1.8: Permuteringstest med $\alpha = 0.05$. Also run with less simulations but same result at 1 mil and 10k sims. The Benjamin-Hochberg procedure on the upper (or lower) triangle reveals that none of the correlations are significant.

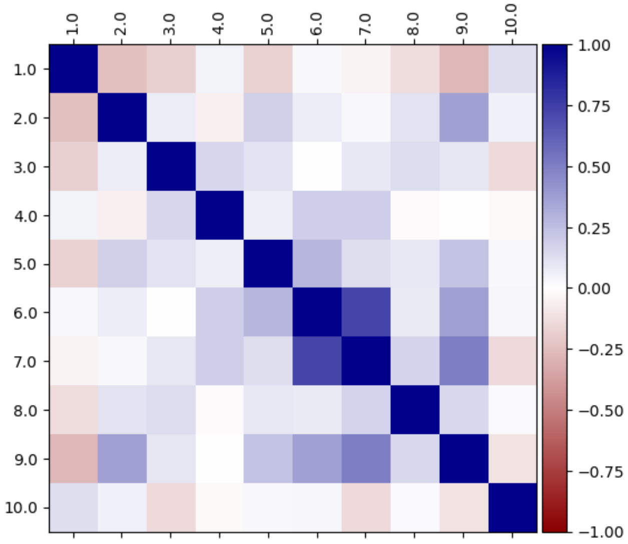

Figure 1.9: Correlation matrix for all phases collapsed

Ligeledes scatter plots for superdiagonalen i ovenstående matrix. Altså phase 1 overfor phase 2, phase 2 overfor phase 3 osv. Er farvelagt efter hvilken cycle de kommer fra. Table 1.4 forklarer hvorfor nogle af de horisontale fremkommer sammen med Figure 1.6 (selve produktionstiden er ret kort sammenlignet med delay.)

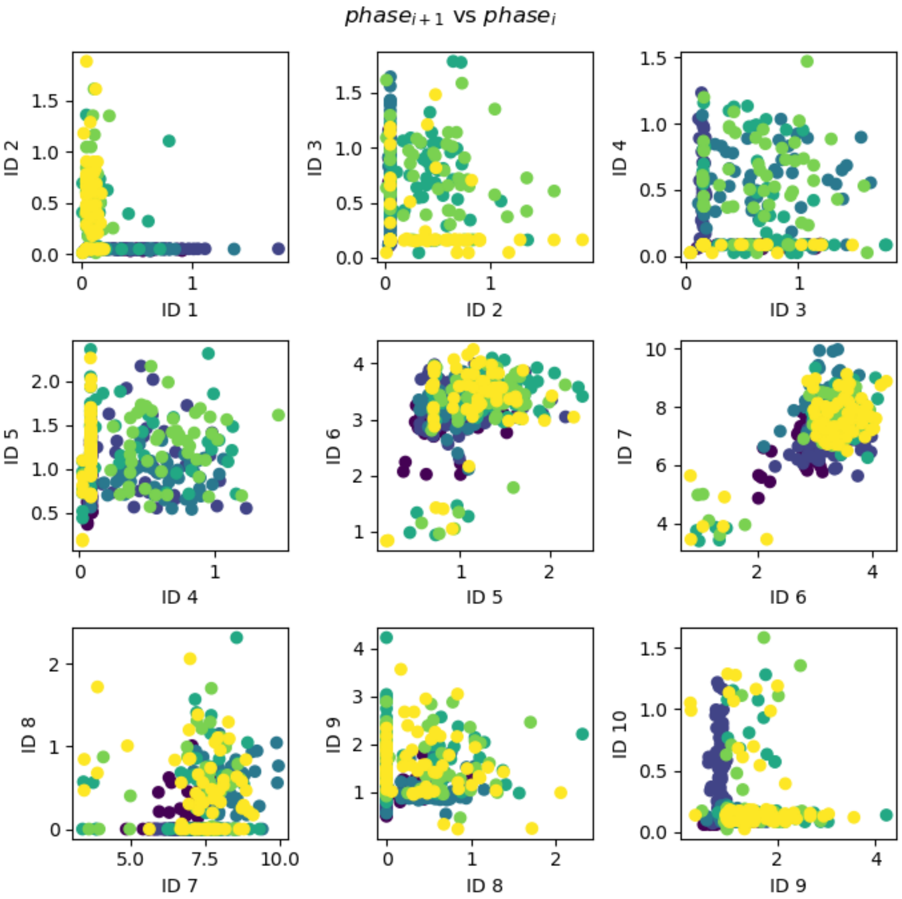

Figure 1.10: Phases vs their next phase when collecting everything regarding a single phase into a total time duration

1.4 Cleaning operations

Sometimes, the vessel is cleansed. This is however not every time after a batch so might be interesting to investigate further. Initially, per cycle, the cleanings are summarized in the following table with basic statistics. As can be seen, there is quite some differences.

The most notifiable differences per batch are the number of cleanses especially when comparing to Table 1.1. For the first two cycles, the cleanses seem to be in between every batch, which is indeed also the while the later four are only sometimes. Furthermore, although the cleanses are between every batch for cycles A and B, the variances are extremely different. For the last four cycles, they seem to be grouped further, E and F are very alike while cleanses in C and D are generally longer although D has a substantially smaller variance than C.

Cycle	#ops	min	max	μ	σ^2	σ	σ/μ
A	65	1.113	3.067	1.917	0.269	0.518	0.270
B	63	1.324	1.751	1.566	0.00883	0.0939	0.0600
C	9	1.544	3.306	2.153	0.277	0.526	0.245
D	10	1.474	2.009	1.581	0.0212	0.146	0.0922
E	10	0.827	1.584	1.465	0.0462	0.215	0.147
F	10	0.748	1.610	1.466	0.0595	0.244	0.166

Table 1.6: Per cycle cleansing statistics

To verify these observations and potentially discovering more important facts of their probability distributions, histograms are plotted in the following Figure 1.11. We indeed again observe the likeliness between the cycles A and B, C and D, E and F respectively. Also, for the first two cycles and more so cycle B, the cleaning times are somewhat normally distributed although cycle A has a very heavy right tail in that case. The later four cycles only have 10 observations but the mode (i.e. peak) seem to be about the same.

From the above observation of like modes one may want to observe the histogram of the combined set of cleaning times. In particular, under the hypothesis that the durations are actually from the same probability distributions and realized independently within each cycle a histogram of all the observations are of interest and is shown in Figure 1.12 below.

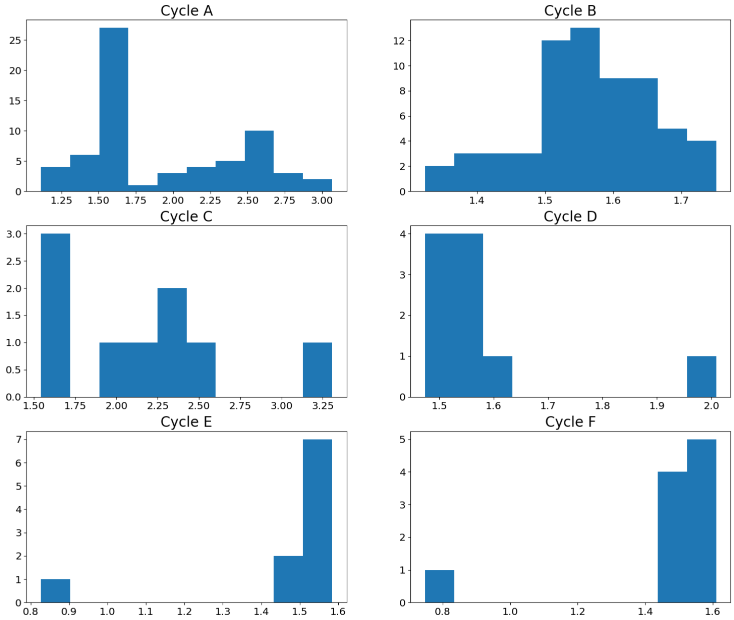

Figure 1.11: Each of the 6 cycles, cleaning operations histograms.

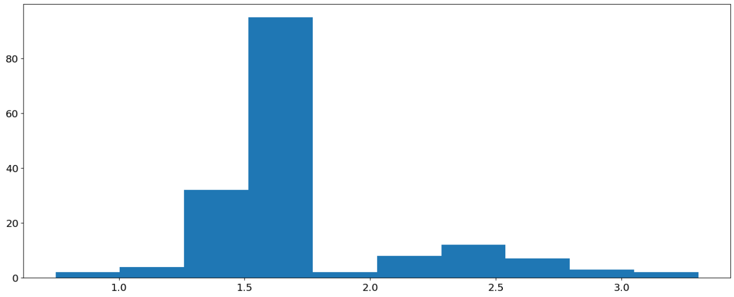

Figure 1.12: Combined cleaning operations histograms.

Finally, to get a better overview of the irregularities is the number of cleaning periods (mostly concerning cycles C through F), each cleaning operation is shown in the following Figure 1.13. The vertical shaded rectangles signify the period in which a cleaning operation is taking place. Furthermore, the event IDs are shown but to get a clearer view on what is going on, a single rectangle (zoomed in) is shown in Figure 1.14.

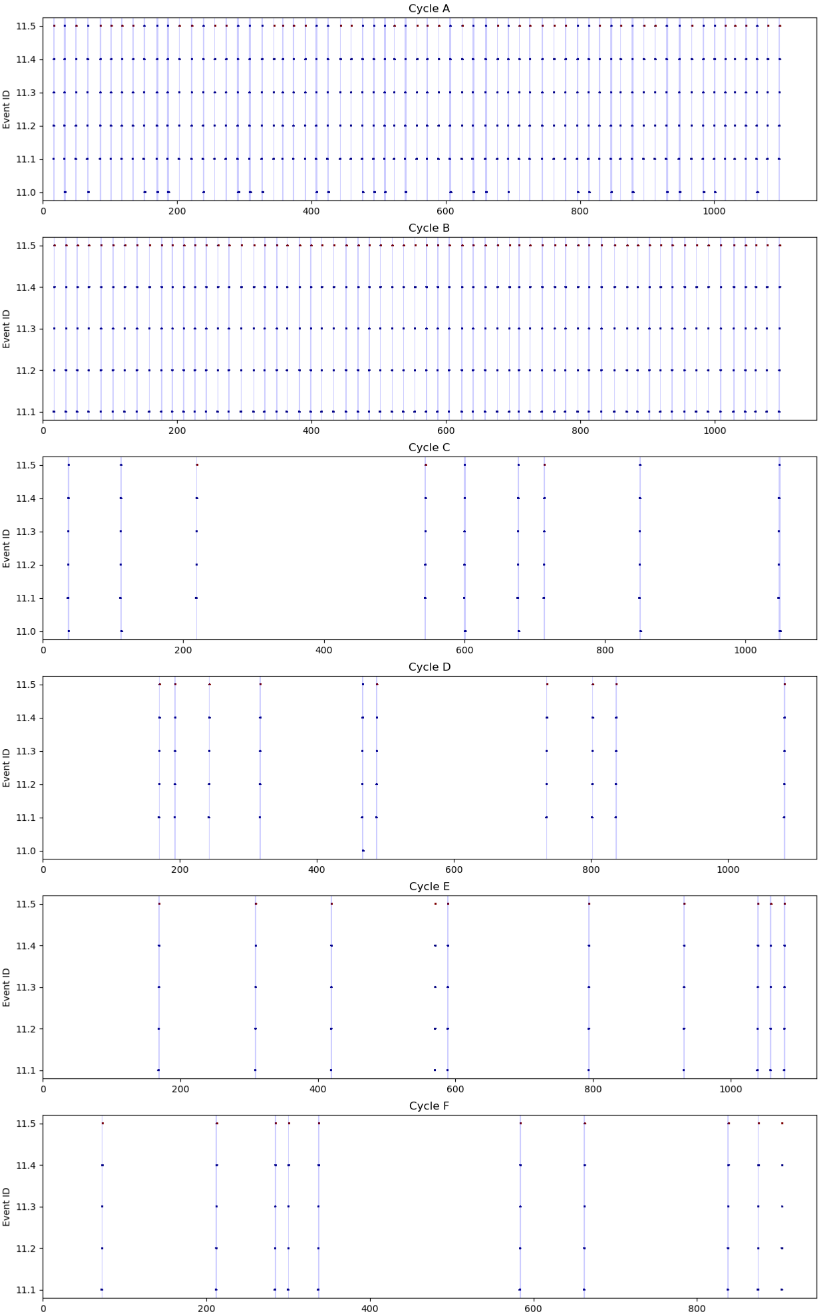

Figure 1.13: Each of the 6 cycles, cleaning (corresponding to `BatchID = 0`). Each *(*Cleaning Procedure), CIP, is highlighted with an opaque interval (the blue rectangles). The dots marked with red (only ID 11.5, but not all of these are red), is if the Cleaning ID is 0.

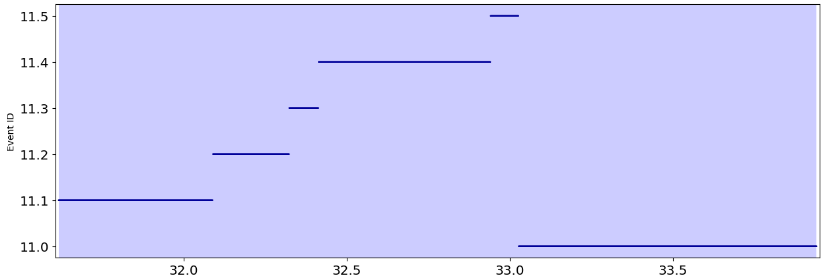

Figure 1.14: A single blue rectangle zoomed in

It is observed that the observations marked with red in figure 1.13 occur exactly when that specific cleaning operation does not go to the state 11.0 after the flush of the tank (event ID 11.5) and vice versa. It is hard to conclude what this may mean, but the cleaning being in state 11.0 may indicate that the system is idle before continuing the next batch like what is observed from the other steps of the process flow. Also, it is noted that while the red dots occur nothing else is happening according to the dataset.

From a modelling point of view, the cycles C through F can be thought of as the cleansing operation having a probability of not happening or equivalently as having a duration of 0. It is thus of interest to observe what the probability of cleaning after an operation is. From Table 1.1 and Table 1.6, we that indeed for cycles A and B, the probability is 100 % when disregarding the possibility of cleaning after the final batch. Hence, we see that for the remaining cycles, the probabilities of cleaning the tank after an operation are as in Table 1.7

Cycle	% cleaning
A	100.00
B	100.00
C	15.00
D	16.95
E	16.95
F	16.13

Table 1.7: Per cycle probability of cleaning

Furthermore, let C_i denote whether the ith batch is followed by a cleaning of the tank or not. It is then of interest if the next batch is followed by a cleaning given whether the current batch is followed by a cleaning. In particular, we count for each of the cycles the transitions which are shown in the following tables. Notice that the number of observations is two less than the total number of batches within each specific cycle. This is due to the last batch is never followed by a cleaning (nor is the first batch superseded by a cleaning procedure) which results in one less observation and also due to the fact that we are logging transitions and hence lose another observation. To test for randomness, a Chi-squared test is carried out on each of the cycles to check for independence. It is observed all the cycles exhibit independence between the groups i.e. there is no statistical evidence for information is gained about if the next batch is followed by a cleaning operation given whether the current batch is followed by a cleaning operation.

C_{i+1} C_i	No	Yes
No	41	9
Yes	9	0

(a) C, $p = 0.3293$

C_{i+1} C_i	No	Yes
No	41	8
Yes	7	2

(b) D, $p = 0.6456$

C_{i+1} C_i	No	Yes
No	41	7
Yes	7	3

(c) E, $p = 0.3532$

C_{i+1} C_i	No	Yes
No	41	9
Yes	9	1

(d) F, $p = 1.0000$

Table 1.8: Contingency table for Cycle C-F

Thus collecting the observations from all the last four cycles, we may want to model the atom of the cleaning procedure independently of the previous batch and with a probability of 0.8375 corresponding to the cleaning procedure only being carried out $16,25\%$ of cases.

Finally, we show the autocorrelation function for each the four cycles C-F in Figure 1.15 and note that all the ACF stay within the 95% confidence interval.

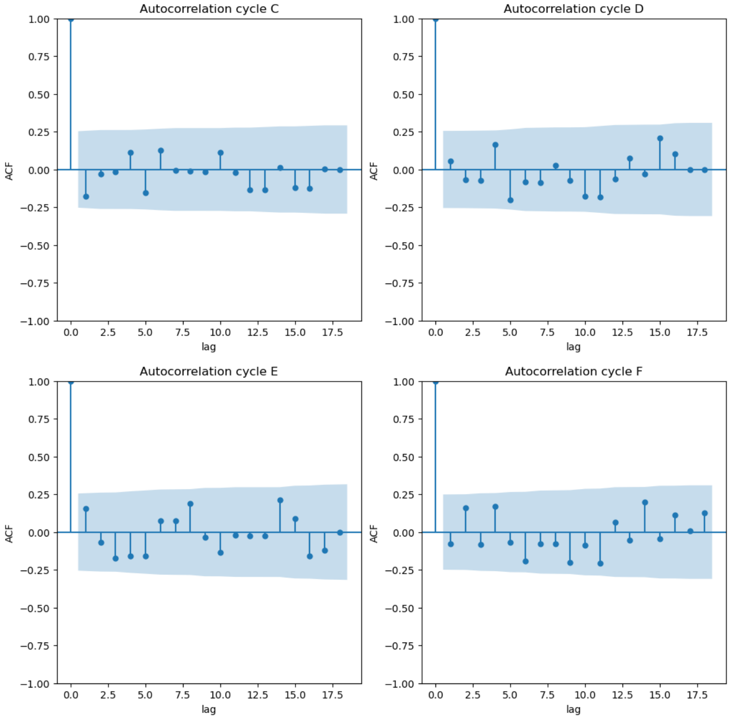

Figure 1.15: Autocorrelation function for each of the final 4 cycles. As can also be seen from this there seem to be no information to be gained of C_i from C_{i-1}.