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.

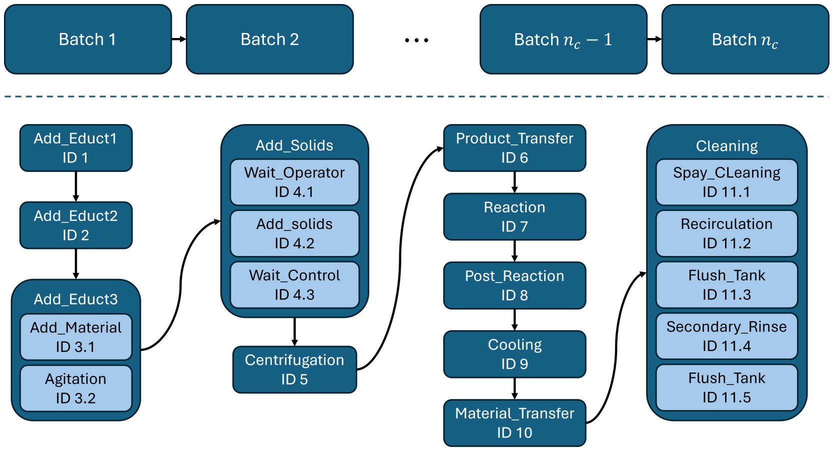

Figure 1.1: The structure of a single cycle. A cycle comprises n_c batches that are carried out one by one. Thus, a cycle is simply a collection of a complex task (a batch), that need to be repeated n_c times. The number of times is often based on time or amount of produced drug or substance, such that 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).

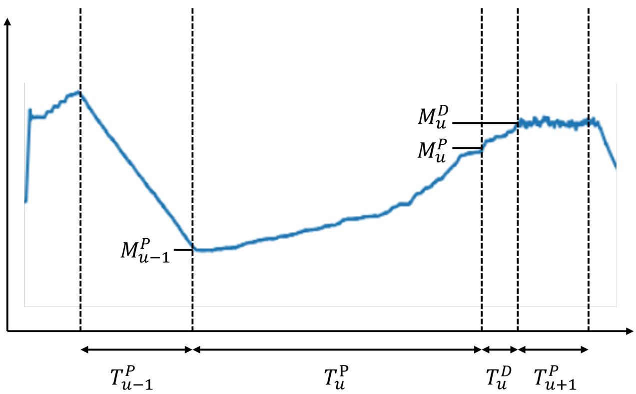

Figure 1.2: Exemplification of the evolution of the level in a tank during a process u and the previous process. The variables T_u^P, T_u^D, M_u^P and M_u^D related to the process are shown. Note that M_u^P and M_u^D are the changes in level from the previous process or delay of process such that they 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

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.

labels pertaining the cleaning operation as these will be handled separately in

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)

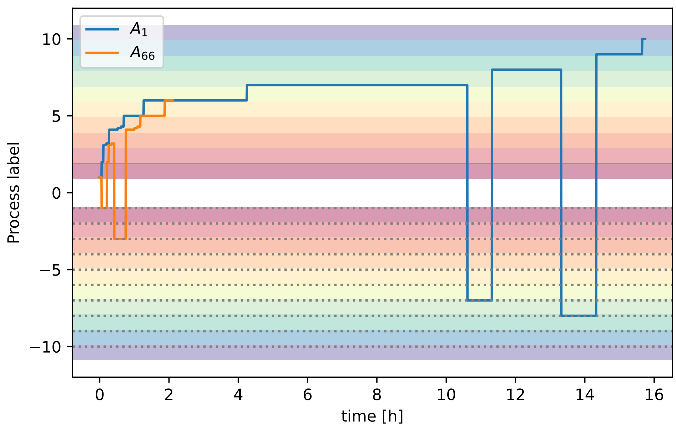

Figure 1.3: The first and last batch from cycle A. The horizontal colored bars corresponds to the different process labels such that time stamps labeled e.g. 3.1 and 3.2 fall in the same colored region. It is clear that the final batch is cut-off before finishing the last process. Furthermore, we observe 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.

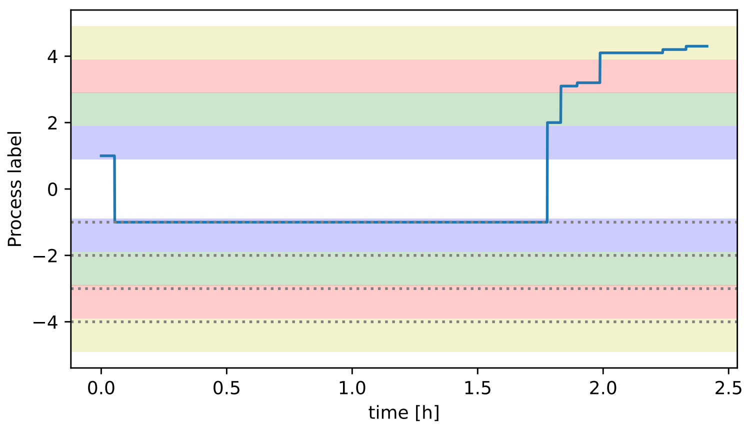

Figure 1.4: The temporal evolution of process labels for batch 22 from cycle B. Only the processes pertaining to the first boxes of Figure 1.1 are shown to easier tell what is happening.

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

Event \ Cycle	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.

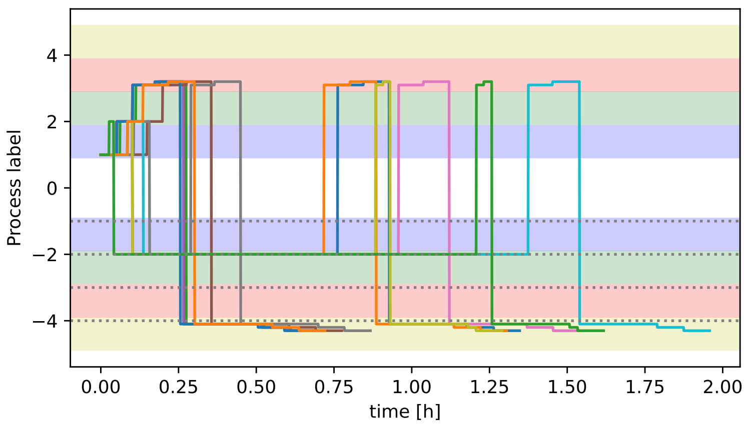

Figure 1.5: 13 out of the total 48 batches where at least one of the process labels -4.1, -4.2 or -4.3 were observed.

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, we finally have a sufficient understanding of the simulation to define a few random variables. For each main process $u \in \{1, \ldots, 10\}$, we define the *delay* after the process as T_u^D and the associated change in level M_u^D. Similarly, for the actual processes $u \in \{1, 2, 3.1, 3.2, 4.1, \ldots, 10\}$ we have *process* duration T_u^P and likewise M_u^P. Converting the time series data to realizations of the random variables, we find that $T_{4.1}^P$, $T_{4.3}^P$ and T_8^P are always constant. Referring once again to [?], we see that indeed these processes are controlled such that the duration is 15 min, 5 min and 2 hours respectively. As these random variables are then not really random but constant, we exclude them from our analysis from this point onward. Note that the delays T_u^D have an atom at 0 since there is a possibility that there is no delay between processes.

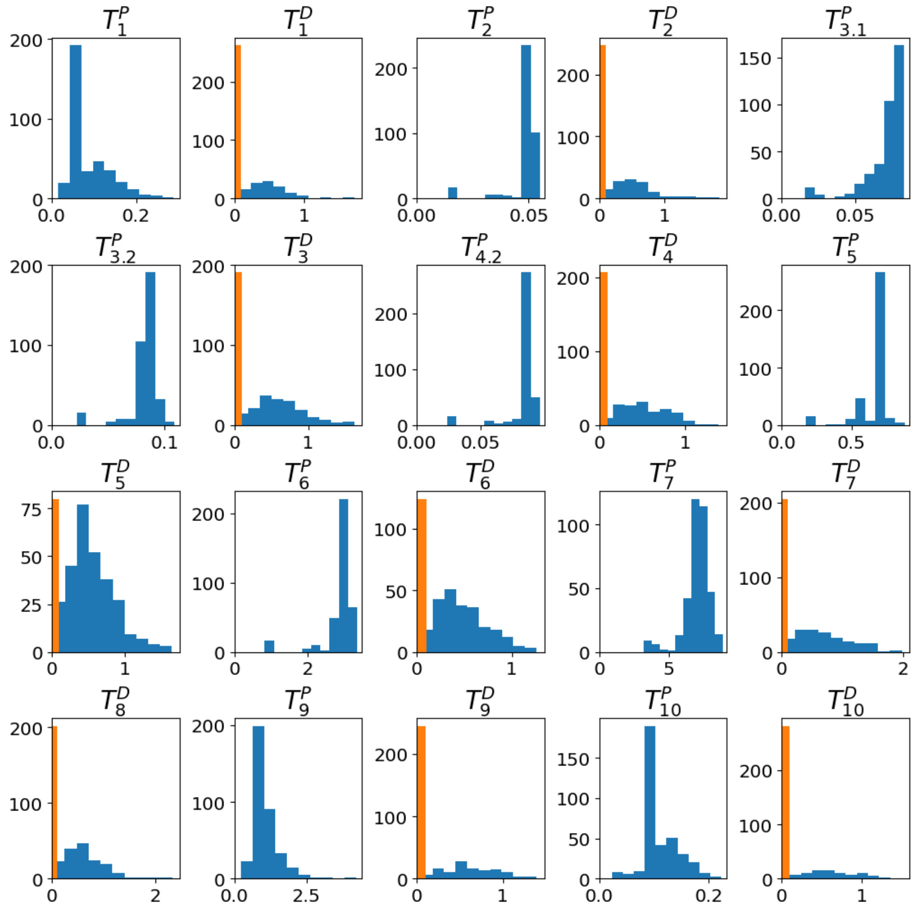

Figure 1.6: Histograms of all random variables T_u^D and T_u^P that are non-constant i.e. not controlled to be a fixed amount of time. The orange bars for T_u^D signify the occasions where no delay after process u took place. We observe that depending on the process, a delay is more or less common. In particular after process 5, there seem to a delay often whereas a delay of process 10 is very rare.

In Figure 1.6, we have shown histograms of frequencies for each of these random variables. We note that for the distribution of observations appear to be unimodal i.e. it does not appear as if they are a mixture of distributions. This is important, as it further strengthens our assumption that the hyperparameters of the simulation where the same for all cycles. In particular, the delays (when disregarding the atom at 0) do not appear as if they originated from different distributions as then we would likely have observed clusters for each of the cycles.

In the next section, we shall further examine the random variables in terms of the correlation structure of the observations. However, we first present some basic statistics of the random variables in Table 1.5 for each cycle. From the

Cycle	A	B	C	D	E	F
$\widehat{\mathbb{E}\left[\sum T_u^P\right]}$	13.993	13.898	15.343	14.471	14.589	14.418
$\widehat{\mathrm{Var}\left(\sum T_u^P\right)}$	0.95636	0.46587	0.76111	4.9589	4.2678	5.3545
$\sum \widehat{\mathrm{Var}\left(T_u^P\right)}$	0.50590	0.31182	0.36667	1.8322	1.5788	1.9696
$\widehat{\mathbb{E}\left[\sum T_u^D\right]}$	0.96398	1.9402	2.4503	3.6050	3.7390	3.0041
$\widehat{\mathrm{Var}\left(\sum T_u^D\right)}$	0.31843	0.39117	0.90187	1.2468	1.2787	1.0462
$\sum \widehat{\mathrm{Var}\left(T_u^D\right)}$	0.34921	0.53198	0.74914	1.4357	1.2454	1.3099
$\widehat{\mathbb{E}\left[\sum T_u^P + T_u^D\right]}$	14.957	15.838	17.793	18.076	18.328	17.422
$\widehat{\mathrm{Var}\left(\sum T_u^P + T_u^D\right)}$	1.1500	1.1352	1.7983	7.0000	5.7086	5.0103

Table 1.5: Mean and variance overview of each of the time related variables T_u^P and T_u^D. A clear difference of variance is observed between cycles $A - C$ and $D - F$. Furthermore, the durations T_u^D clearly show that the total variation is accounted for by the variance of the individual durations. The contrary holds true for the delays T_u^D.

table, we observe that the average total durations (excluding delays) of batches are approximately the same. However, the variance of the sum of durations is significantly larger for the cycles D, E and F. What causes this difference in variation is however unknown. Ideally, in a real world application, one should investigate this further. On could argue that cycles $A - C$ and $D - F$ should then be treated separately and indeed the variance for the accumulated delay during a batch exhibit the same behavior. Namely, the delays also seem to have a larger variance in the later three cycles. We shall however treat all the batches simultaneously in this thesis by arguing that although there is a clear difference in variation, the underlying causal structure of the processes is assumed to be the same. In particular, we could not infer from the simulation study [?] that the cycles should have been based on different hyperparameters resulting in the below difference of variances. Furthermore, we note that the difference between the variation in the accumulated duration and the sum of process durations i.e. $\widehat{\mathrm{Var}\left(\sum T_u^P\right)} - \sum \widehat{\mathrm{Var}\left(T_u^P\right)}$ indicate that the durations of the processes are not unrelated. Thus, in the next section we shall investigate the correlation structure of the variables. Finally, we note that the same difference for durations does not indicate a relationship. However, the missing covariances might just cancel each other out. Hence, we do not yet conclude anything regarding their causal structure.

1.3.1 Correlation structure of durations and delays

In this section, we proceed with investigating the correlation between pairs of the random variables. Based on the observations, we quickly compute a correlation matrix as observed in Figure 1.7.

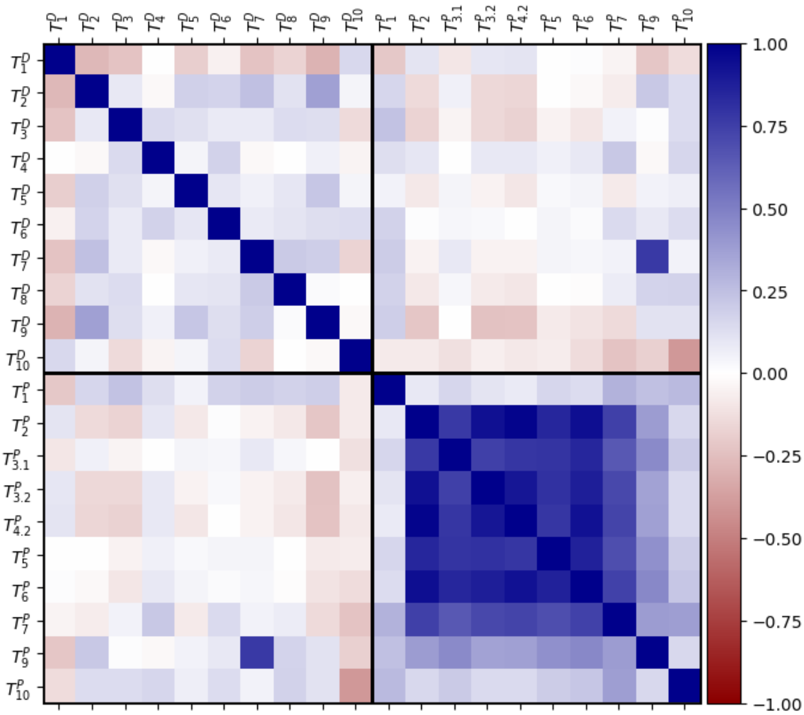

Figure 1.7: Estimated correlation matrix of the random variables T_u^P and T_u^D.

We immediately notice that the durations of the processes appear to be positively correlated. The fact that there exists strong correlations is not surprising when considering Table 1.5 where we observed that the variation of the sum of durations was much larger than the sum of individual variances. The only other immediately interesting observation that we make from the above figure is the large correlation between T_7^D and T_9^P. This means that there is a positive linear relationship between the delay after the reaction process 7 and the time it takes to cool the tank (process 9).

To assess the significance of these correlations, we performed a permutation test. Namely, by randomly permuting the observations of each random variable and recomputing the correlation matrix, we see if the new correlation coefficient is numerically larger than the one we computed from the original data. Repeating this multiple times (e.g. 10,000 times), we obtain an estimate of the probability of observing a correlation coefficient, numerically larger than the one we computed in Figure 1.7. In particular, the null hypothesis is that there is no cor-

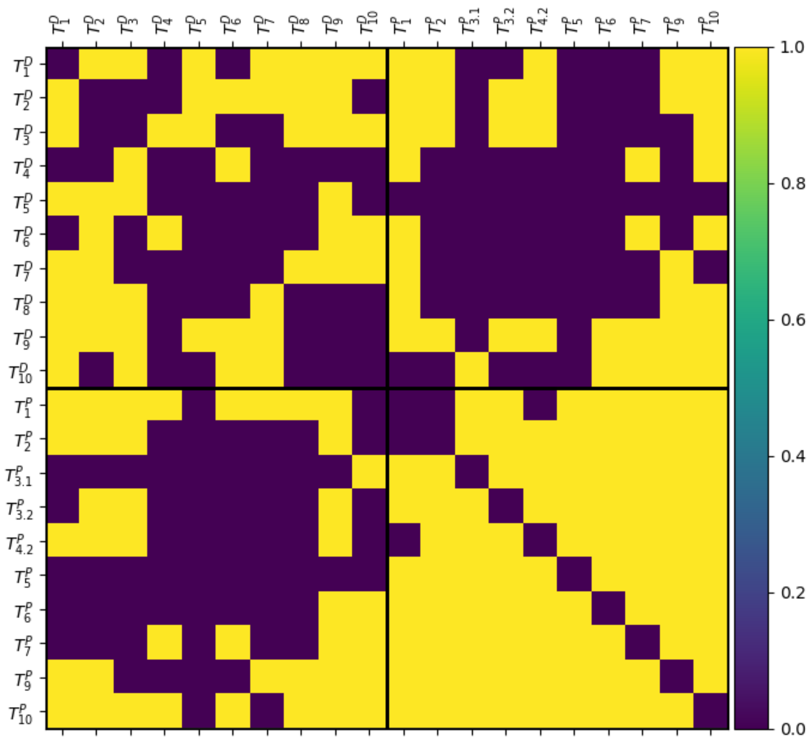

Figure 1.8: Permutation test with significance level $\alpha = 0.05$ based on 10,000 permutations. The diagonal elements have been set to 0 as it does not make sense to test the correlation between a random variable and itself.

relation between a pair of random variables. This is because when the samples are reshuffled independently of each other, the underlying assumption is that the random variables are independent of each other and hence has correlation 0. In Figure 1.8 we have shown a binary matrix for when the p-value was observed to be less than 0.05 i.e. significant results on a 5% significance level. We observe that many of the correlations appear to significant which is somewhat contradictory to what we would expect from such a production system. However, as we are performing multiple test, we really should correct for this in some way. Choosing a conservative approach through the Bonferroni correction, we get far fewer significant results as observed in Figure 1.9. Once again, we have removed the diagonal elements.

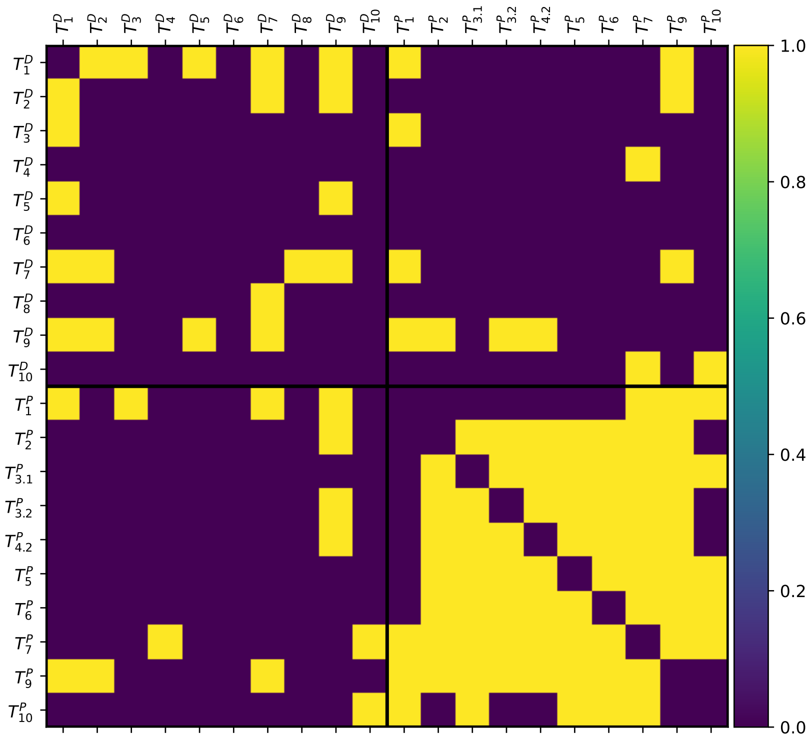

Figure 1.9: The permutation test with Bonferroni corrected significance level. Many of the significant correlations disappear however when plotting e.g. T_1^D vs T_2^D we see that the correlation does not stem from a linear relationship. In particular, from **??**, we observe that the joint distribution is more of an L-shape. We shall thus later investigate other measures of similarity than correlation to capture such non-linear tendencies

Suppose now that we instead did not know what part of a process what the duration and what was a delay. In particular, we define the total duration of a process as $T_u = T_u^P + T_u^D$. Note that for T_3 and T_4, we extend this definition slightly such that $T_3 = T_{3.1}^P + T_{3.2}^P + T_3^D$ and $T_4 = T_{4.1}^P + T_{4.2}^P + T_{4.3}^P + T_4^D$ (although $T_{4.1}^P$ and $T_{4.3}^P$ can be disregarded as they are constant and hence irrelevant when computing the correlation). Using only the cumulated random variables, we observe a much simpler correlation structure as shown in Figure 1.10. However, we also seem to lose much of the information that was otherwise visible in Figure 1.7.

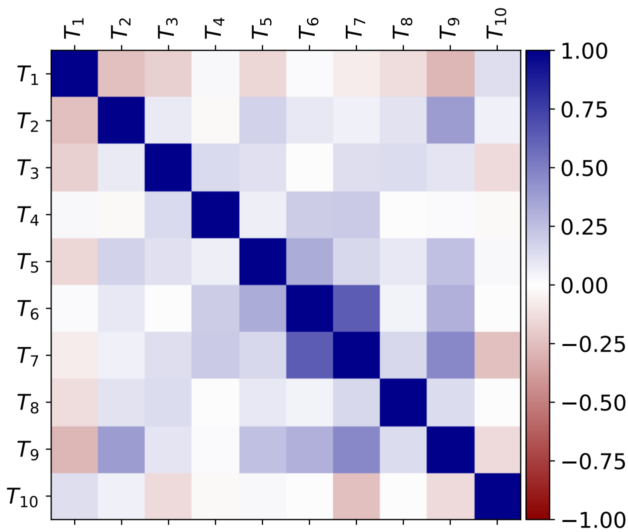

Figure 1.10: Correlation matrix for combined process durations T_u. The strong correlations that we observed previously appear to be lost. In particular, it is no longer clear what durations influence each other if any. Thus, we conclude that at least in the case of linear relations between durations, the extra information regarding when the system is idle due to a delay, is very important if we want to have any change of inferring anything useful about the causal structure of the production system.

As a final remark, we have shown in Figure 1.11 the duration of a process T_u and the process that follows immediately after T_{u+1}. Although the correlations in Figure 1.10 did not appear as informative as in Figure 1.7, there might still be some useful observations to be made from the joint raw observations. Note that Cycle 1 corresponds to Cycle A and so on. From the below figure, it is clear that the duration of a process is not well-described based solely on the previous process. However, we do observe an interesting behavior in T_2 vs T_1 and T_{10}

vs T_9. Namely, for T_2 and T_1, T_1 appear to be larger in the first 3 cycles while T_2 appears to be larger in the final three cycles. This results in the L-shape observed. A similar observation is made between T_{10} and T_9. From Table 1.4, we see that this coincides with the existence of delays in the cycles respectively. Furthermore, from Figure 1.6, the durations often last much longer than the actual duration of the process, such that the delay dominates the total duration of the process.

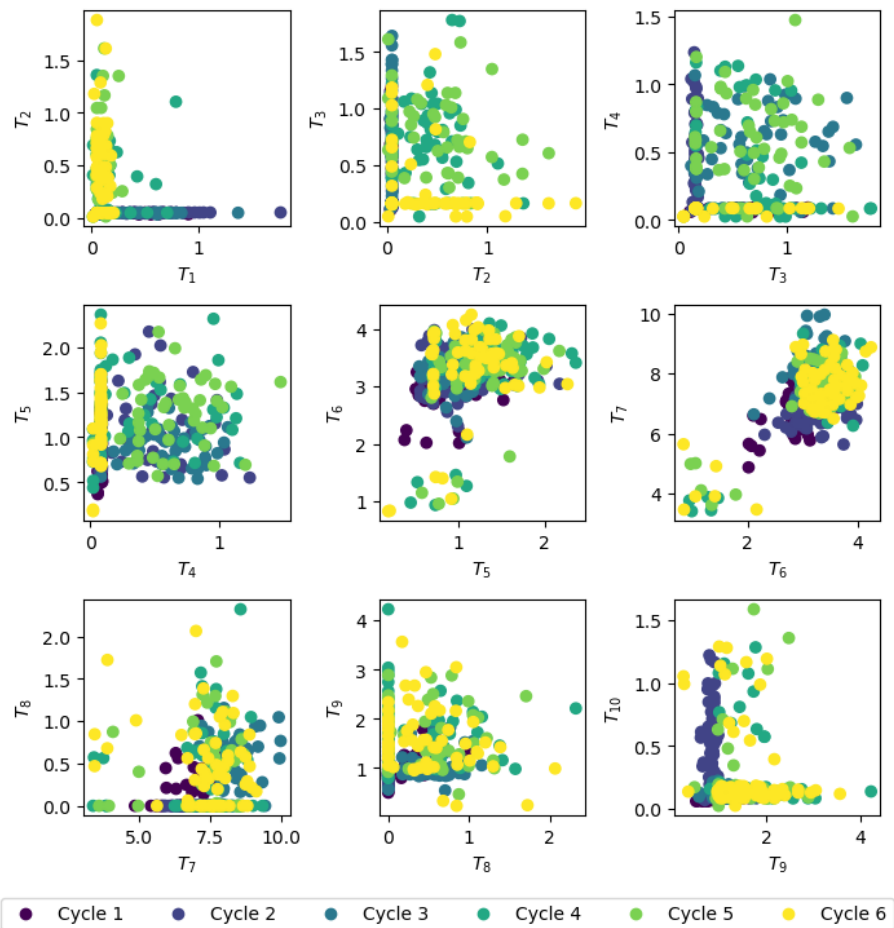

Figure 1.11: Total process durations T_u plotted against the next process T_{u+1}. In some cases, there seem to be a difference from cycle to cycle. Especially in T_2 vs T_1 and T_{10} vs T_9.

For the sake of completeness, we have in the appendix, **??**, **??** and **??** plotted all variables against each other. Although clear relations can be observed between some pairs of random variables it is unclear how e.g. durations and delays (and the related level changes) propagate through the system if they even do so. In **??**, we will discuss a method for discovering such relations, but first, we comment on the cleaning operations that until this point has been left out.

1.4 Cleaning operations

As per Figure 1.1, after each batch, the tank in which the process has taken place is to be cleaned. However, from inspection of the data, we observe that only for cycles A and B is this true. Namely, for cycles C through F, the tank is not cleaned after each batch. This, along with some basic statistics regarding the duration of the cleaning operation is summarized in Table 1.6.

Cycle	number of cleanses	min	max	sample mean	sample variance	sample standard deviation	CV
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. We observe that cycles A and B have many more cleaning processes taking place than cycles $C - F$. Also, the variation in cycles A and C is much larger than for the other cycles. These observations indicate that the cleaning operations from the different cycles have not been simulated from the same hyperparameters. This apparent inequality between the cycles is the primary reason for why we have not chosen to use them in our further analysis of the production system.

The most notable differences from cycle to cycle are the number of cleanses as each cycle have approximately the same number of batches as seen in Table 1.1. For the first two cycles, the cleanses seem to be in between every batch, which is indeed the case. Furthermore, although the cleanses are between every batch for cycles A and B, the variances are extremely different. Also, we observe that cycles A and C have much larger variation than the remaining cycles. With these observations, we hypothesize that the cleaning operations have not been simulated based on the same settings whence we will not consider these processes in our further analysis.

Furthermore, histograms of the durations of cleaning processes for each cycle in Figure 1.12 show that although cycle B has many more cleanses, the distribution is somewhat comparable to those of cycles $D-F$ as the durations are approximately centered around 1.55 with similar interquartile range.

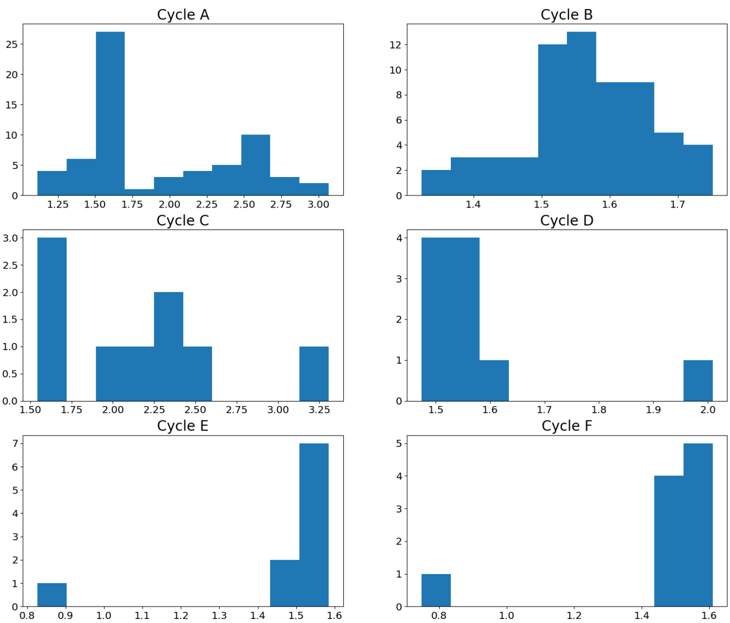

Figure 1.12: Each of the 6 cycles, cleaning operations histograms.

In the appendix, **??** and **??**, we have shown how a cleaning operation goes through each of the subprocesses. Indeed, for cycles $C-F$ the cleaning process is only carried out every so often as summarized in Table 1.7. From **??**, the cleaning processes appear to be carried out between a random number of batches. In the remaining part of this section, we shall investigate this observation for the latter four cycles with some simple tests.

In particular, let C_i denote the indicator of whether the ith batch is followed by a cleaning of the tank or not. We shall then investigate whether the next batch is also cleaned or not. I.e. at a lag 1, if the variables C_i are associated. In particular, we shall use Fisher's exact test with the alternative hypothesis being two-sided. We use the `scipy` implementation `stats.fisher_exact`. The results are summarized in Table 1.8. Indeed, we observe that the results are non-significant on a 5% significance level. Repeating the tests for lags up to and including 10 we observe no significant results either. More sophisticated

Cycle	Percentage cleaning processes after batches
A	100.00
B	100.00
C	15.00
D	16.95
E	16.95
F	16.13

Table 1.7: Per cycle, how often a batch is followed by a cleaning process.

tests could be carried out to test if using e.g. the previous 5 C_i is predictive of whether the next batch is followed by a cleaning process. However, as we shall not use this later, we end our discussion of the cleaning processes and note that in the remaining of the thesis, they have been excluded from the dataset.

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

(a) Cycle C, $p = 0.3293$

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

(b) Cycle D, $p = 0.6456$

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

(c) Cycle E, $p = 0.3532$

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

(d) Cycle F, $p = 1.0000$

Table 1.8: Contingency table for Cycle C-F at lag 1. No significant results are observed and repeating for lags larger than 1 we do not conclude otherwise.