Fault Detection and Sensor Diagnostics Using Control Charts

Group Members

1. Aime Barema
2. Nadine Bisanukuli Cyizere
3. Socrates Waka Onyando

Introduction

We have been provided with a dataset related to wind turbine failures. The observations were recorded at 10-second intervals through sensors within the Supervisory Control and Data Acquisition (SCADA) system. While the exact variable names remain undisclosed, they correspond to diverse physical process properties. The dataset encompasses data from four distinct wind turbines, denoted as `WT2, WT3, WT14`, and `WT39`.

Among these turbines, we have been informed that `WT2` is indicative of a healthy, functioning turbine. In contrast, the other three turbines exhibit various stages of faults, which initially manifest as faults and subsequently transition to normal operation once the fault-related issues have been rectified.

Modelling Goal

This project focuses on the utilisation of control charts and sensor diagnostics for fault detection in wind turbines. Control charts, created based on a healthy turbine model, can be used to identify faults in the faulty turbines. The project aims to optimise the selection of principal components within the healthy model and subsequently generate control charts using multivariate statistical techniques, specifically T^2 and SPEx control charts. Through a detailed analysis of the contributions to individual out-of-control observations, the project seeks to pinpoint the sensors that effectively capture and diagnose turbine faults.

Communication Channels

For this project, we will utilise several communication and collaboration tools. WhatsApp will serve as our messaging platform, Google Meet will be employed for virtual meetings, and GitHub will facilitate code sharing. Additionally, we will schedule in-person meetings as necessary to address project needs.

Data Importation

The data has been accurately imported into an expected format using MATLAB's `readtable()` function, ensuring that observations are represented as rows and variables are structured as columns.

Data Description

As already mentioned, the wind turbine dataset comprises measurements from a total of four turbines. Details regarding the number of observations and features are presented in Table 1 for reference.

Wind Turbine	Number of Observations	No of Features
WT2	1570	28
WT3	698	31
WT14	686	27
WT39	1405	27

Table 1: *Number of observations and features per wind turbine*

It's evident that the datasets for WT2 and WT3 contain a greater number of features compared to the datasets for WT14 and WT39. Moreover, measurements are taken at 10-second intervals from the wind turbine sensors, resulting in a dataset that constitutes a time series of measurements.

Summary of Data

The summary of our data provides a concise overview of its key characteristics. It encompasses information on the central tendency, dispersion, and distribution of our dataset. Some of the key elements of the summary for each turbine are presented in Table 1 (WT2), Table 2 (WT3), Table 3 (WT14), Table 4 (WT39). Notably, the data variables have no labels linking them to any physical features. Therefore, it is difficult to know what the variables mean.

Variable	Mean	Std dev.	Max	Min	Median
1	-0.007	0.046	0.118	-0.161	-0.011
2	-0.012	0.074	0.249	-0.239	-0.011
3	26.312	1.462	27.700	23.300	26.900
4	29.280	1.632	31.000	25.748	29.800
5	315.310	117.136	516.923	96.323	316.207
6	41.044	2.289	43.598	36.498	41.798
7	40.906	2.239	43.496	36.496	41.596
8	41.080	2.233	43.531	36.595	41.795
9	23574828.594	397.961	23575361.000	23574057.000	23574923.000
10	320.622	117.091	523.190	103.614	321.193
11	1320.342	146.661	1555.433	1061.598	1326.353
12	1750.000	0.000	1750.000	1750.000	1750.000
13	0.011	0.014	0.484	0.010	0.010
14	-0.754	1.597	4.000	-6.000	-1.000
15	849.000	0.000	849.000	849.000	849.000
16	223.684	15.407	263.697	200.828	227.835
17	9.043	1.441	11.900	6.974	8.900

18	685.091	2.149	697.852	678.076	685.268
19	50.002	0.045	50.080	49.910	50.020
20	17.623	0.302	17.974	16.900	17.800
21	22.912	0.286	23.202	22.200	23.039
22	10.736	0.220	11.100	10.400	10.700
23	5.164	4.674	18.364	-10.767	5.222
24	5.480	0.784	7.370	3.405	5.632
25	16.923	2.563	24.607	9.209	16.795
26	53.211	3.145	60.003	49.604	51.504
27	59.822	1.900	63.954	56.000	59.500
28	60.626	2.590	65.596	55.097	59.996

Table 2: *Summary statistics for variables of* `WT2`

Variable 12 stands out since it has a standard deviation of 0. This suggests that it may be constant. Variable 9 also has high values as compared to the rest.

Variable	Mean	Std dev.	Max	Min	Median
1	0.000	0.000	0.000	0.000	0.000
2	49.986	0.030	50.060	49.930	49.980
3	0.044	0.000	0.044	0.044	0.044
4	13.457	10.426	45.669	-9.162	13.156
5	12.315	1.875	28.000	12.000	12.000
6	67.309	126.479	359.424	0.797	16.588
7	88.993	0.001	88.993	88.969	88.993
8	0.000	0.000	0.000	0.000	0.000
9	50.000	0.000	50.000	50.000	50.000
10	89.000	0.000	89.000	89.000	89.000
11	-0.010	0.031	0.246	-0.268	-0.011
12	-0.010	0.018	0.089	-0.078	-0.011
13	0.018	0.022	0.235	0.000	0.011
14	0.015	0.007	0.067	0.000	0.011
15	0.420	2.004	6.101	-4.437	0.201
16	4.061	1.515	7.395	0.470	4.195
17	2.814	1.725	7.196	-0.925	2.496
18	42.452	1.340	46.727	40.100	42.500
19	849.000	0.000	849.000	849.000	849.000
20	41.722	6.089	59.189	33.490	40.577
21	24.338	2.912	28.700	19.600	24.207
22	27.727	4.667	35.201	20.301	27.383
23	38.240	2.837	42.897	33.297	38.286
24	38.849	2.613	43.267	34.294	38.888
25	38.913	2.402	42.942	34.697	38.897
26	-0.153	0.643	1.999	-1.001	-0.401
27	-0.326	0.825	0.542	-4.400	-0.200
28	849.000	0.000	849.000	849.000	849.000
29	7.742	0.449	8.597	6.997	7.697
30	13.719	1.118	15.600	11.600	13.800

31	-5.774	0.797	-0.828	-6.501	-6.101

Table 3: *Summary statistics for variables of* `WT3`

For `WT3`, several variables stick out. This is because they have 0 standard deviation. This includes variable 28, variable 19, variable 8, variable 9, variable 10, variable 1 and variable 3.

Variable	Mean	Std dev.	Max	Min	Median
1	-0.009	0.041	0.170	-0.256	-0.011
2	-0.008	0.064	0.249	-0.239	-0.011
3	50.848	22.981	76.798	8.226	67.498
4	51.691	21.275	76.299	6.742	67.348
5	201.143	215.402	1021.722	0.022	0.033
6	59.396	17.525	88.994	41.098	62.979
7	59.242	17.585	89.271	40.996	62.770
8	59.316	17.490	89.299	41.149	62.604
9			25990129.000	23574057.000	
10	203.845	219.203	1256.034	0.000	0.000
11	722.773	953.523	17052.875	-19.697	23.887
12	862.755	849.878	1750.000	0.000	50.000
13	46.237	44.416	88.990	0.010	88.989
14	-0.264	2.945	68.000	-6.000	0.000
15	847.762	32.415	849.000	0.000	849.000
16	101.540	97.210	204.695	3.307	13.276
17	24.521	12.969	40.532	2.299	35.199
18	686.923	26.086	696.654	13.475	686.466
19	49.944	1.910	50.090	0.002	50.040
20	27.251	9.122	36.600	0.101	35.700
21	32.135	8.870	42.300	0.711	40.000
22	22.576	11.128	35.000	0.280	32.800
23	2.074	8.047	80.968	-21.251	2.989
24	6.967	1.194	11.031	1.015	6.661
25	35.398	15.906	57.796	13.379	45.740
26	63.853	6.024	77.495	23.298	61.395
27	64.984	6.681	84.566	41.207	61.096

Table 4: *Summary statistics for variables of* `WT14`

For `WT14`, variable 9 is the obvious stand out because it has a missing value.

Variable	Mean	Std dev.	Max	Min	Median
1	-0.009	0.043	0.192	-0.192	-0.011
2	-0.010	0.067	0.249	-0.239	-0.011
3	28.644	2.072	32.700	1.116	27.700
4	31.592	2.208	36.652	3.089	30.800
5	266.596	196.408	516.923	-0.019	335.770
6	44.256	2.743	50.800	2.747	43.298
7	44.217	2.883	50.798	2.404	43.096

8	44.250	2.697	50.199	1.921	43.268
9	22361582.965	1792517.322	23575045.000	423.644	23574305.000
10	270.117	198.809	523.190	0.000	341.462
11	962.621	660.099	1555.433	0.011	1351.744
12	1181.281	802.435	1750.000	0.000	1750.000
13	29.711	41.973	88.991	0.010	0.010
14	-0.522	1.389	4.000	-6.000	0.000
15	568.315	396.033	849.000	0.241	849.000
16	240.062	39.369	294.631	3.465	219.841
17	11.059	1.770	14.500	0.963	10.800
18	685.222	18.355	697.852	1.950	685.867
19	49.962	1.335	50.080	0.001	50.000
20	20.238	3.466	25.900	0.241	17.900
21	24.500	2.117	28.100	0.202	23.148
22	12.861	2.948	20.700	1.236	11.100
23	3.944	4.817	19.071	-21.840	4.009
24	6.884	1.393	10.472	0.936	6.297
25	14.099	5.165	21.995	0.049	16.695
26	63.932	7.211	80.698	11.515	59.600
27	63.845	7.201	88.298	49.377	60.596

***Table 5**: Summary statistics for variables of `WT39`*

Visualisation of Dataset

From Figure 1 below, we can see that the dataset from *`WT3`* is the obvious standout in terms of variable distribution. Moreover, this dataset stands out scalewise. As earlier pointed out, variable 9 has very high values as compared to the other variables. This explains why the boxplots for variables in *`WT2, WT14`* and *`WT39`* have very large scales (up to 10^7).

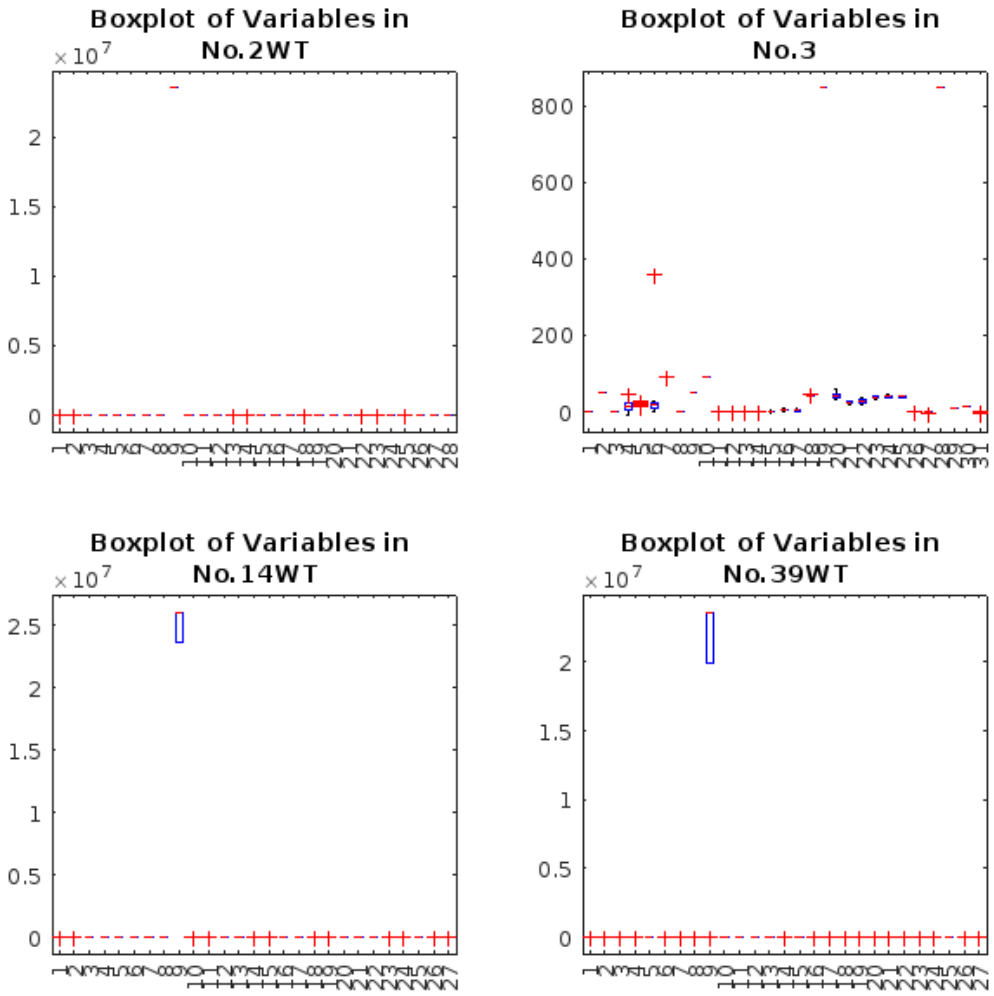

Figure 1: Boxplots of variables from the turbines.

Challenges of the Data

- Absence of labels for the data features: This hinders our ability to understand the features, analyse them correctly in their proper variable types (continuous, categorical etc) and contextual some of the results from our model.
- Different number of features: As shown in Table 1, different numbers of measurements were collected from each turbine.
- No clarity on time series synchronisation: Whilst we have time series data, it is not clear whether the observations were collected synchronously. This is further evidenced by the fact that the number of observations per turbine are not the same.
- Missing data. We have identified a single missing value for the turbine `WT14`.

Data Pretreatment

- Discard the dataset for *`WT3`* given the above outlined issues.

- Drop one of the columns in the dataset for `WT2` in order to get an equal number of features for each of the datasets used in the analysis. This is because the dataset for `WT2` has one more feature than the other two datasets.
- Determine a strategy for handling the missing value.
- Centre and scale the datasets
- Perform feature selection and/or engineering