2. Data

2.1. Data sources

Dataset used in this thesis is provided by Ericsson site in Linköping, Sweden. Ericsson founded by Lars Magnus Ericsson since 1876 is one of the world's leading in telecommunication industry. It provides services, software products and infrastructure related to information and communications technology (ICT). Its head quarter is located in Stockholm, Sweden. Ericsson continuously expands its services and products beyond telecoms sector such as mobile broadband, cloud services, transportation and network design.

The dataset is collected on 20 January 2017 and is extracted from log file produced by test cases. There are different types of test cases which are being executed in the simulation system. Each test case is viewed as an observation in the dataset.

2.2. Data description

This is the dataset for the second generation (G2) product which contain 2,xxx test cases. Following are the variables in the dataset:

Metadata of test case

- Timestamp: Date and time when a test case is being executed (yy-dd-mm hh:mm:ss)
- NodeName: IP address or the name of base station
- DuProdName: Product name
- Fdd/Tdd: Different standard of LTE 4G Technology. Fdd and Tdd stand for frequency division duplex and time division duplex, respectively.
- NumCells: Number of cells in the Antenna
- Release: Software release
- SW: Software package for the software release
- LogFilePath: Path for log file produced by a test case

Observable memory

- MemFreeKiB:
- SwapFreeKiB:
- BufferCacheKiB:
- PageCacheKiB:
- RealFreeKiB: True free memory

CPU

- TotCpu%: Total or average CPU usage. This value represents the performance of the test case in terms of CPU utilization.
- PerCpu%: CPU usage of each CPU
- PerThread%:
- EventsPerSec: Events or traffic intensity

  This variable contains several components that can be used when defining the test case. Apparently, there is no fixed number of components in this variable as different test cases involve additional procedure compared to other test cases. The components along with their values in EventsPerSec field are also varied depending on which types of test cases are being executed. These components and their values are the main factor which have an impact on the CPU utilization. The unit for each component is its event per second.

Some variables in the dataset are chosen to be used for further analysis. There are in total one response variable and six predictor variables. Table 2.1 shows the name of variables and their description followed by the type and unit measure. The first three predictor variables are components of the test case which can be found in the EventsPerSec field while the last three variables are considered as the test environment. These variables appear to have a high contribution to the CPU utilization.

Table 2.1.: Description of the selected variables

Variable	Name	Type	Unit
Response	TotCpu%	Continuous	Percentage
Predictor	RrcConnectionSetupComplete	Continuous	Per second
	Paging	Continuous	Per second
	X2HandoverRequest	Continuous	Per second
	DuProdName	Categorical	
	Fdd/Tdd	Binary	
	NumCells	Categorical	

There are three software releases.

2.3. Data preprocessing

The dataset is sorted by the version of the software package which is named alphabetically. Even though the dataset has a timestamp when the test case is run, the timestamp is unsuitable to represent itself as a time point in time series. The software package is an important variable and more of an interest to this analysis. It is, therefore, seen as the data points indexed in time order in the time series.

Test cases are not always being executed properly. It is either no traffic is generated during the test case or that data is not logged. If the EventsPerSec field is missing or has no value in it, the test case is being treated as incomplete and all the data related to that particular test case is ignored. This also applied for other cases when any fields in the dataset is missing.

There are some columns in the dataset that store multiple values separated by a tab character. These tab-separated values are split to columns in order to use for later analysis. This process is done for the EventsPerSec variable in order to get its components and values which characterized the test case.

Each software release consists of several software packages. Also, numerous test cases are executed in one software package. Since the software package acts as a point in time in the time series, it appears to be rather difficult to visualize all results from every executed test case for each software package. Hence, the test case which has the lowest value of the CPU utilization (or minimum value of TotCpu%) is selected to represent the performance of a specific software package.

Although taking an average of the multiple runs for test cases in the software package appears to be a good approach, it does not yield the best outcome in this case. The first reason is that manipulating data can easily be misleading. It is, therefore, settled to remain the data as it is and always visualize the data exactly as it is recorded. Another important reason for not using the average value of the CPU utilization is because the essential information in the test case will be lost. Each test case has its own components in EventsPerSec field that used for identifying the test case. The details of these components are absent when averaging over the CPU utilization of test cases in the software package.