The aim of this work is to structure and combine the results of geotechnical ground drilling done with the CPT (Cone Prediction Test) and the associated laboratories testing. The combined results will constitute a dataset intended to be used in a machine learning model to predict soil types on sample wells before being sent to the laboratory. There will be scripts created in Python that automate creating the combined dataset, it is also important that the dataset is designed to be adapted for use in a machine learning algorithm. The total number of samples is over 10,000.

Data from the base drills are stored in a so-called .snd file and the lab reports in Excel format. The important data in the .snd files is; coordinates (1, 2), altitude (3), depth (12), pressure (13), torque (14) and flush pressure (15). From the lab reports, data on soil type will be obtained mainly. This means that when multiple datasets from ground drilling and lab reports are to be combined, that dataset will consist of several dimensions because each test has different data at different depths.

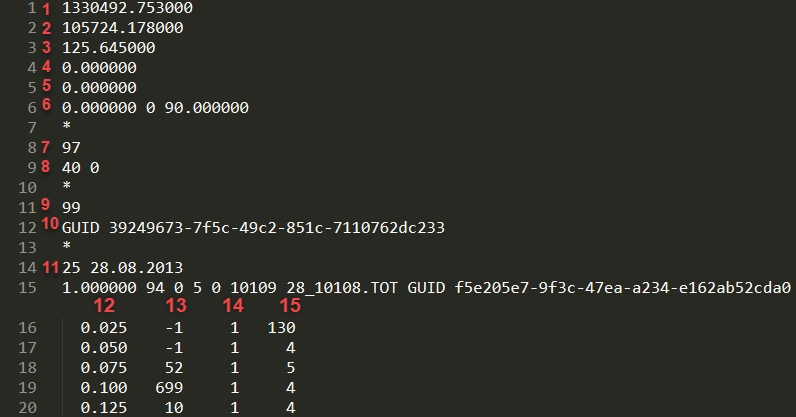


Figure 1 - .snd base drilling (CPT) file  
1: Coordinate  
2: Coordinate  
3: Height above sea level  
10: Unique identifier  
11: Date  
12: Depth (m)  
13: Pressure (kN)  
14: Torque (Sec \* 10)  
15: Flushing pressure (kN)

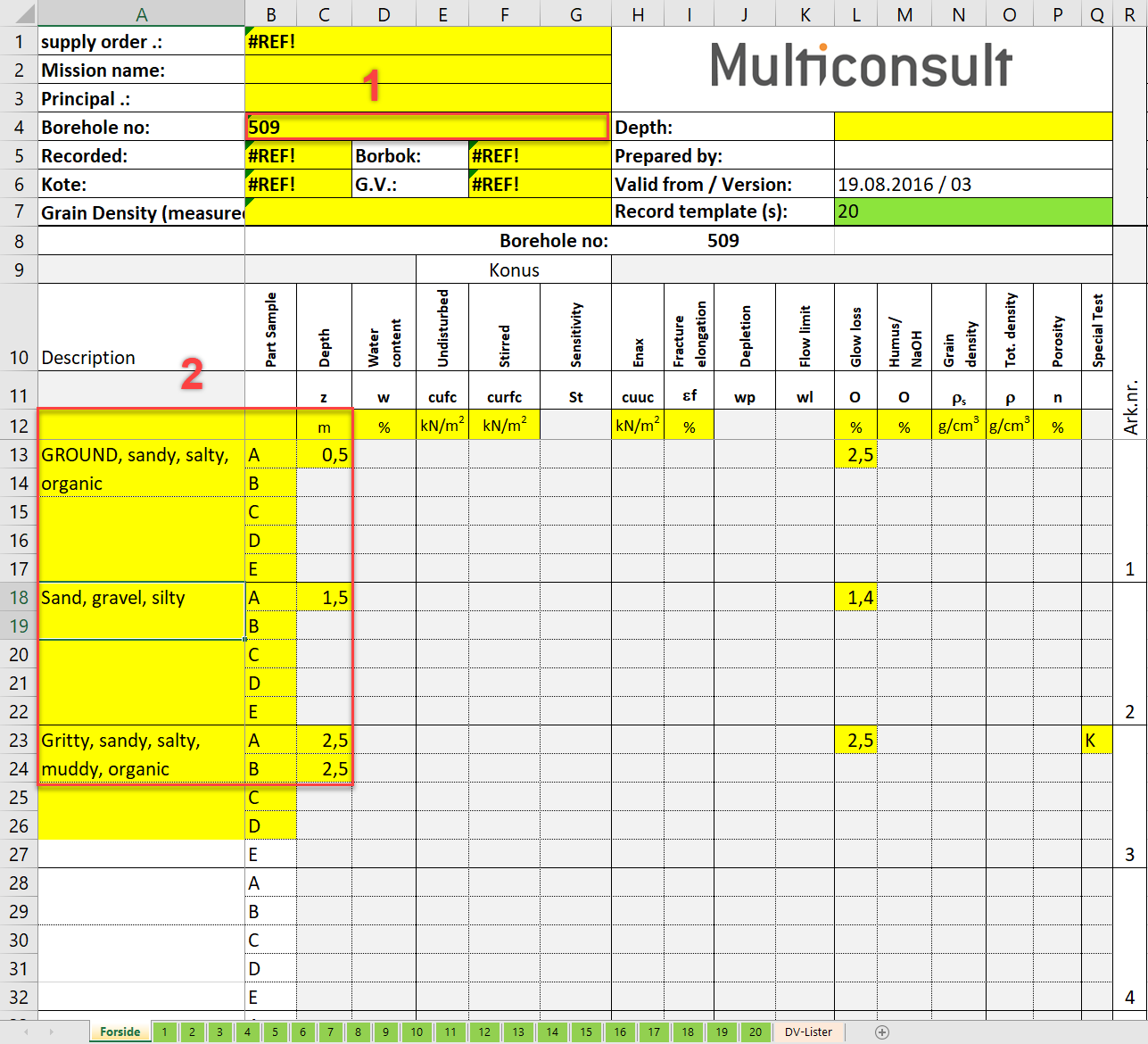


Figure 2: Laboratories report  
1: Reference to borehole and .snd file  
2: Soil description with depth