Yujia Zhou
Dickinson College
Mentors: Emery Boose and Barbara Lerner

Quality Control of Sensor Data and Data Provenance Tracking

Scientists often rely on sensors to obtain data. Sometimes, sensors may go wrong, thus the raw data needs to be processed before it can be used. This process of quality control includes but is not limited to calibration adjustment, detection of irregular values, and gap filling of missing data. In this research, we used the 15-minute real-time data from the meteorological station at the Harvard Forest and studied the quality control methods that might be applied to this dataset. We developed R programs to detect and fix quality control problems. This process will be performed multiple times in the future due to improvements of quality control techniques and hence will generate different versions of datasets. However, as the datasets grow larger and time passes, it becomes difficult to know how a particular version of the dataset was derived from the raw data, because we will lose track of which activities have been done to obtain a specific version of the processed data. As a result, recording the data provenance, or the history of data, is necessary for scientists to understand data derivation. Data Derivation Graphs (DDG) can record the full provenance of how each data point is derived from the raw data, allowing scientists to keep track of their data. To accomplish this goal, we built a process simulating scientists' initial processing of the raw data and the reprocessing after some of the quality control techniques are updated. We implemented this process in both Kepler and Little-JIL to compare the data provenance graphs they produce from identical processes. We found that while Kepler is easier to use for scientists with no programming background, Little-JIL has a much stronger visualization tool for drawing comprehensive DDGs and stores more information in them.