University of BRISTOL

School of Physics

Name:	Wesley Richardson
Degree Course:	Physics with Scientific Computing BsC
Student Number:	U 2164839

Assessed Exercise – Assessment of Topic 1: Scientific Data Analysis

Option 2: Hubble Constant

Abstract:

Hubble’s constant describes the rate of cosmic expansion, and can be calculated from the velocity and distance of a distant galactic object. There has been a significant deviation from the accepted Theoretical value of Hubble’s constant to the recorded literature value. Over time, this deviation has decreased, as scientific apparatus, understanding and methods have improved. The latest experimental models are very close to the theoretical value. This report showcases one such example where the value deviates from the ideal and explores why and how.

This report investigates the value of Hubble's Constant (H0) values over time, utilizing a comprehensive dataset comprising various measurements from multiple sources. An ideal value of H0 was extracted from the Hubble 2 dataset by analysing data to find strategical filters to inspect the highest quality data with the greatest accuracy and precision while still maintaining a signifcant proportion of data points.

Statistical analysis and linear regression models were applied to extrapolate an ideal H0 value from a filtered subset of the Hubble 1 data set. Further graphical and statistical analysis was performed to investigate fluctuations in the value of H0 within Hubble set 1. Deviations from the linear regression model were used to construct residual plots which demonstrated that deviations are proportional to the distance squared and appear suddenly as the data training range is exited.

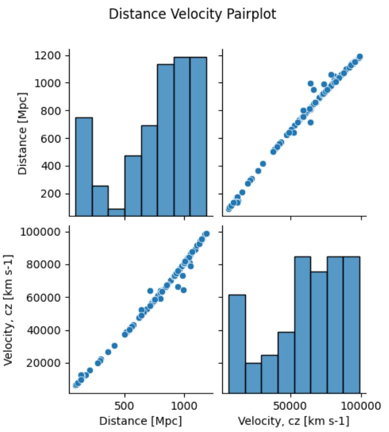
Finally the models constructed from Hubble 1 data were compared to ideal data stratified from Hubble 2 data. The ideal value of Hubble’s constant derived from Hubble 2 was taken as the mean and was found to be 68.69km/s/Mpc with a standard deviation of 7.42km/s/Mpc. This deviated from that of Hubble 1 data which was found to be 71.8km/s/Mpc with a standard deviation of 1.2km/s/Mpc. A scientific critique of the Hubble 1 data ensued to understand the disagreement.

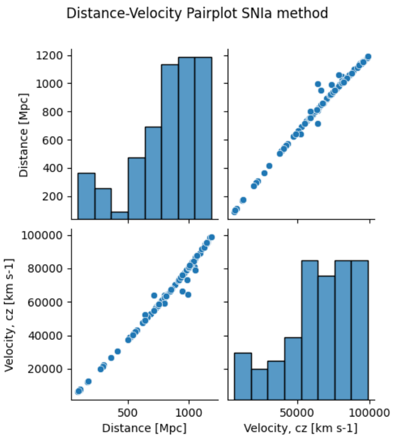
Part a: Creation, assessment and comparison of two suggested linear fit models.

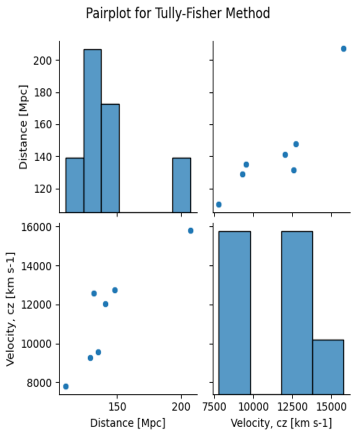
Within ‘Ex_hubble1.csv’ (Hubble 1 csv) there exist 101 records containing experimental observations on the distance [Mpc], velocity [km/s] and velocity uncertainty [km/s] of astronomical objects. There is also a Method column, which states the method used to gather the data.

The Hubble 1 data set was investigated. It was found using the pandas.df.corr() method that the object velocity column was correlated with the object distance with a correlation coefficient of 0.995, demonstrating a high likeliehood of Hubble’s relation.

Using Seaborn, velocity distance pairplots were created to visually demonstrate their relationship. There are two distinct method values in the Hubble 1 methods field; ‘SNIa’ and ‘Tully-Fisher’. Two filtered dataframes were created containing only records of a specific method. Pairplots were created to check if either of these methods is superior.

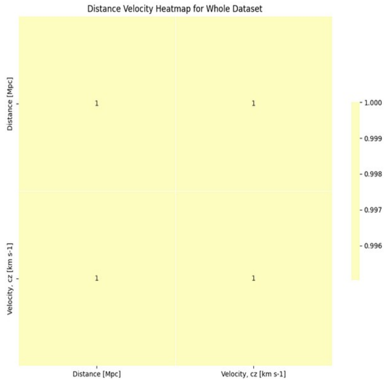

Distance Velocity Pairplot

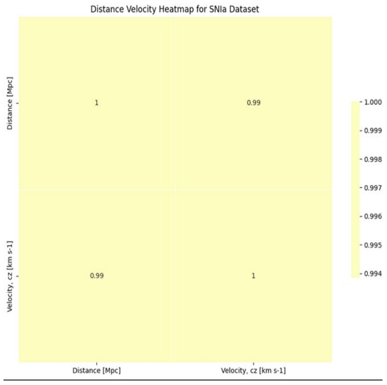

Distance-Velocity Pairplot SNIa method

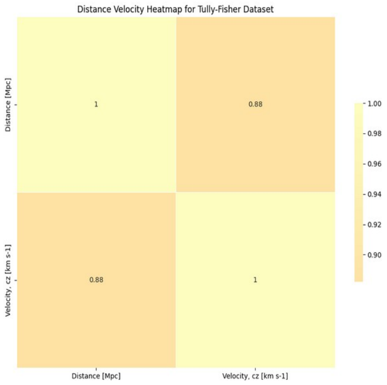

Pairplot for Tully-Fisher Method

Within the SNIa plot there are slow changes in the gradient of the linear relationship which begin at about 500 [Mpc].

The Tully-Fisher data represents only a small portion of the total data points. Additonally it seems to demonstrate a weaker linear relation with data points not following a clear gradient. For more precise investigation, a heatmap will provide numerical and visual insight into the correlations.


Distance Velocity Heatmap for Whole Dataset

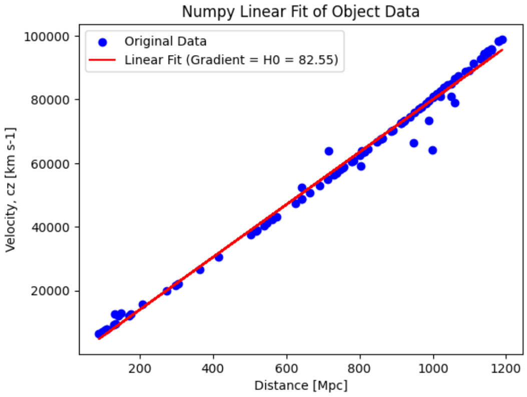

Distance Velocity Heatmap for SNIa Dataset


Distance Velocity Heatmap for Tully-Fisher Dataset

Interestingly, depsite the SNIa dataset visually following a stronger linear fit, after isolating it from the Tully-Fisher data, the correlation coefficient has decreased. However, the Tully-Fisher dataset does present with a much weaker coeffficient of 0.88. This suggests that error in the Tully-Fisher set is offsetting any non linear behaviour from the SNIa data set.

Tully-Fisher data seems to have a lower quality of data for the purpose of finding an ideal H0 value, since it seems to contains more linear deviation and there is a significantly greater quantity of data within the SNIa method.

A numpy 1st order linear fit was implemented as a preliminary investigation into the value of H

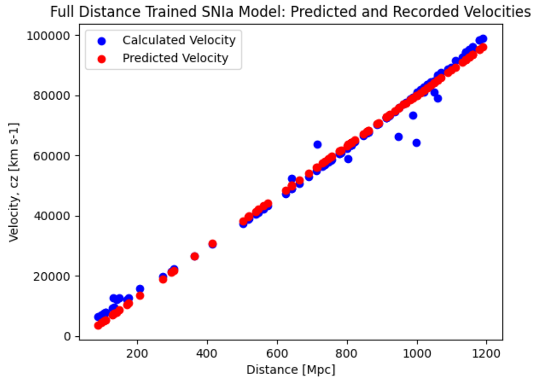

Numpy Linear Fit of Object Data

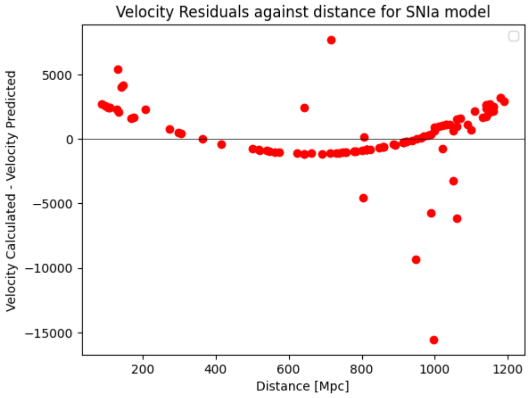
An estimated H0 value of 82.55 km/s/Mpc can be taken as the gradient of the line.

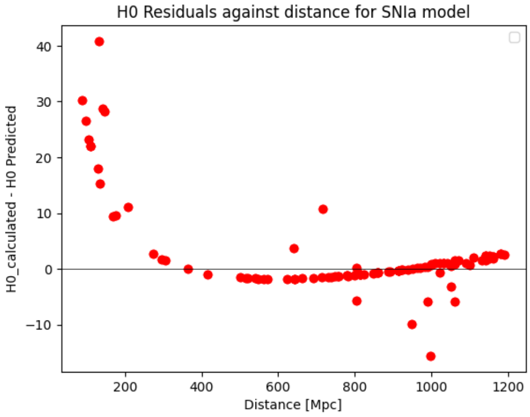
A more suitable value of H0 can be taken as the mean calculated H0 value of the dataset. The error may be taken as the standard deviation.

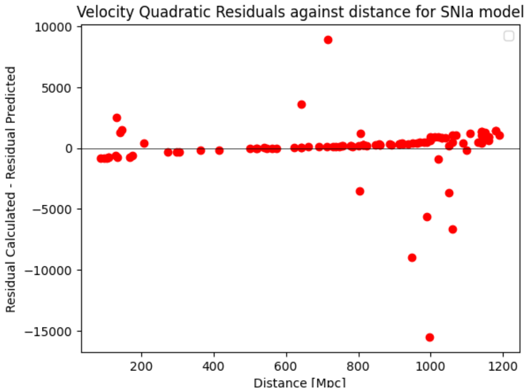
Linear regression models will yield a more elegant fit since its function is to minimize the residual sum of squares between provided data and predicted data.

The first linear regression was trained using the whole range of distance-velocity data. An additional plot was created to investigate how the residuals of calculated and recorded velocities vary with distance
For completeness, an additional plot was created showing H0 residuals over the distance.









The value of Hubble’s constant attained as the slope of the linear regression model is 83.90km/s/Mpc which is less accurate than the value attained using the numpy simple linear fit.

The mean value of H0 was found to be 78.18 km/s/Mpc with a standard deviation of 3.98 km/s/Mpc. The median value of Hubble’s constant was found to be 78.7 km/s/Mpc.

The velocity residual plot appears to demonstrate quadratic behaviour. A numpy order 2 polyfit was applied to the velocity residuals as a function of distance. The residuals residuals of the predicted and actual residuals were also plotted, the meta-residuals appear to follow a linear trend.

This data suggests that Hubble’s constant is not constant. This effect may be a result of cosmic dust, which can propagate errors into distance and velocity calculations.

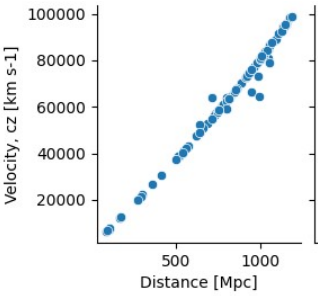
Cosmic dust will absorb light emitted from astronommical objects, leading to a decreased intensity on the sensor. This obstruction causes inaccuries in the methods used to determine object distance from incident light.

Furthermore, dust particles will also scatter light, leading to a reddening of light. This change in color of light will effect the measurement of both the distance and velocities of distant planets.

The combination of these two effects introduces systematic errors in the estimation of the diameter and luminositiy of distant astrolonomical bodies, these errors propagate into the calculations of distance and velocity, which influence calculations ofHubble’s constant.

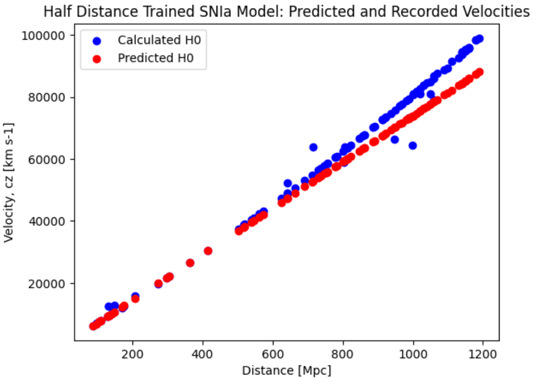
We would expect to see the model predict a more accurate value of Hubble’s constant for the data set which is trained using only the first half of the distance data, since these data are less likely to be influenced by cosmic dust.

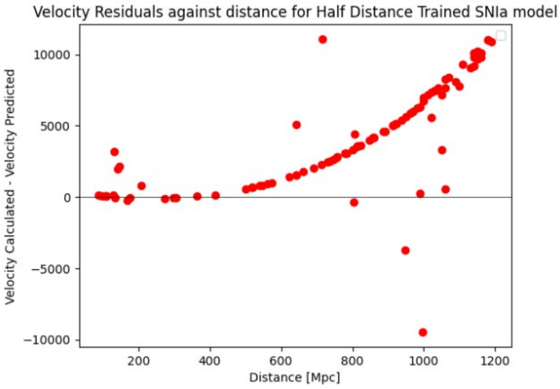
Cosmic dust explains how we see the velocity gradient increase in this plot on the right. The distance of farther objects is more likely to be under-estimated, hence how the gradient decreases at higher distances. Far objects have an unexpectedly higher velocity because their recorded distances are smaller than their actual distances.



Following this, we would expect that a model trained using only low distances would yield a more accurate value of Hubble's constant.

Linear regression model trained with points for distance < 500 Mpc:

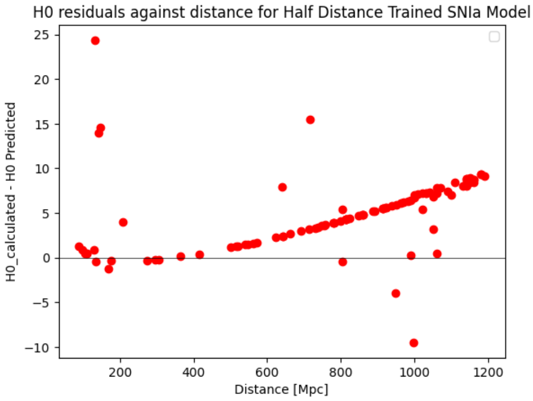




The value of Hubble's constant as the slope of the linear regression fit is 74.36 km/s/Mpc. The mean calculated value of Hubble's constant is 71.82 km/s/Mpc with a standard deviation of 1.2 km/s/Mpc. The median was found to be 71.54 km/s/Mpc.

We can see that the residuals follow a exponential or quadratic trend with distance. This demontrates that the error is more significant for objects at larger distances. This can be attributed to how light emitted from farther objects will interact with more dust on its path to the sensor.

There are also error points that do not follow this trend. This may be explained by astronomical objects which are of very large velocities, overpowering the effect of cosmic expansion. Equally, these points could be novel astronomical objects for which the normal methods of calculating velocity and distance do not work well.



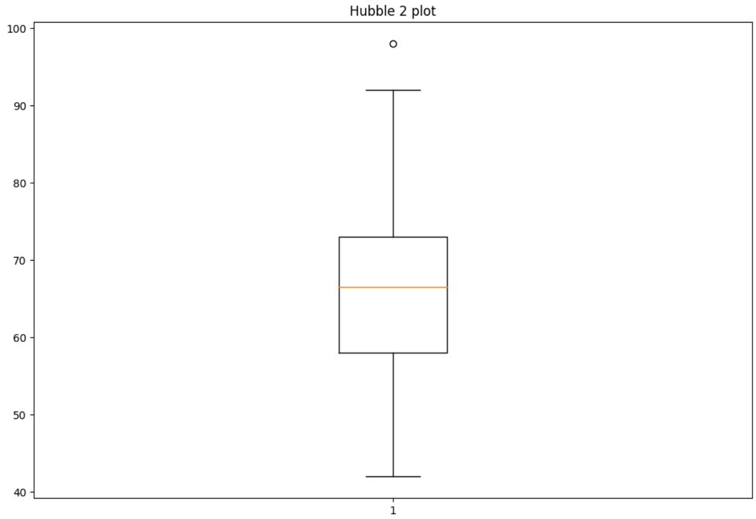
Here we have a graph of H0 residuals against distance, which also purports the cosmic dust explanation to the variability of Hubble's constant.

Assuming that linear discrepancies are caused by cosmic dust, the half SNIa trained data set would offer a more accurate value for Hubble's constant, this value of 71.28km/s/Mpc is also within the accepted literature range.

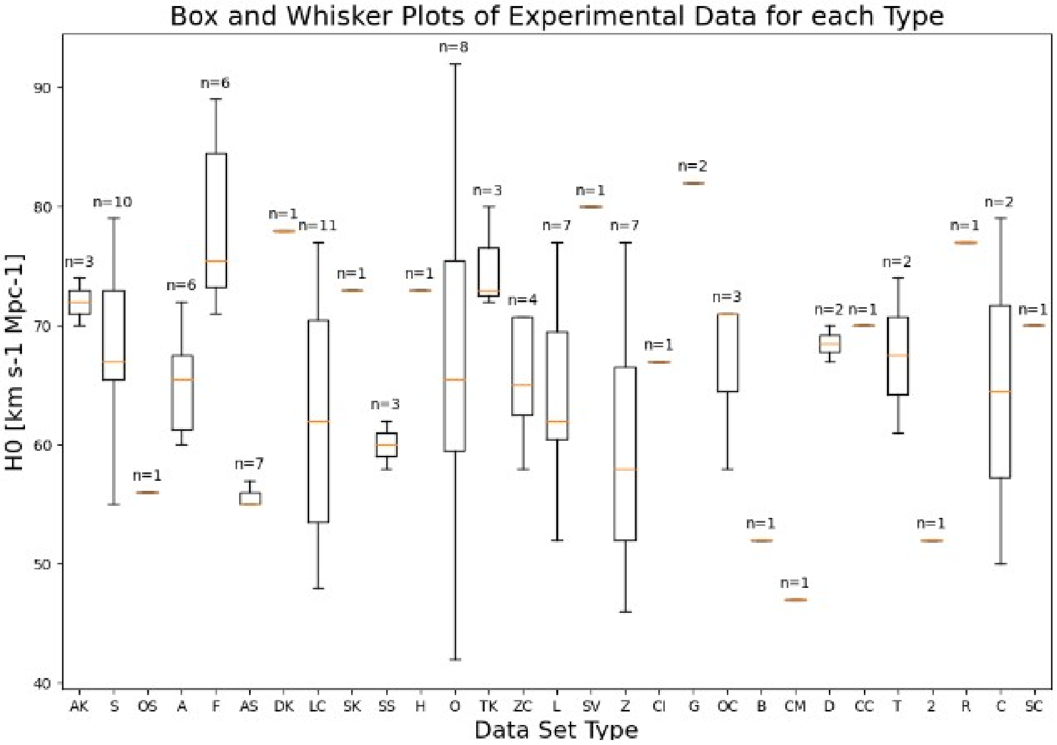
Part b) Use statistical measures of location and variability, visualisation plots, to compare new data to the values for Hubble's constant determined in part a). justify why you have chosen the descriptive statistics used and critically assess the models fitted in part a), comparing them to the "ideal" value of Hubble's constant suggested by the collected dataset.

The data set ‘Ex_Hubble_2.csv’ (Hubble 2 2) contains fields for; Hubble’s Constant, errors (+), errors (-), Date ,experiment ‘Type’ and a source field stating a reference.

Immediately, it is possible to take the mean value of H and the associated error on the mean, yielding H = 66.29 km/s/Mpc with a standard deviation of 11.23km/s/Mpc. The set has a range of 46 km/s/Mpc, and a median of 66.5 km/s/Mpc.

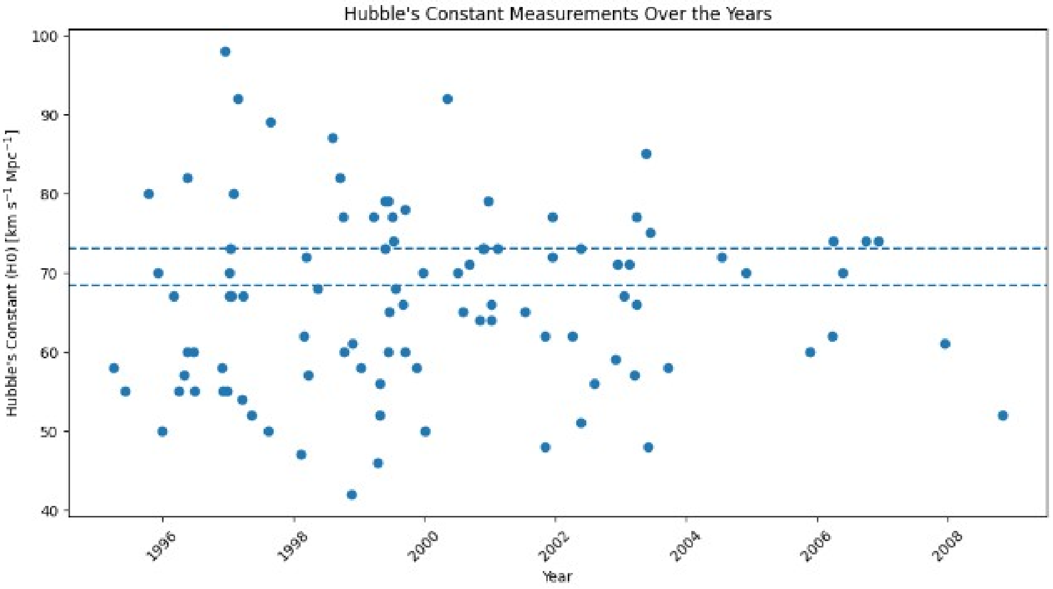


The dataframes were seperated by their type value. A box plot of Hubble’s constant for each type will demonstrate the mean, range, and interquartile range associated with each type. It is possible that there are differences as to the accuracy and precision of Hubble’s constant measurements associated with the type, and a box plot will do well to demonstrate these trends.

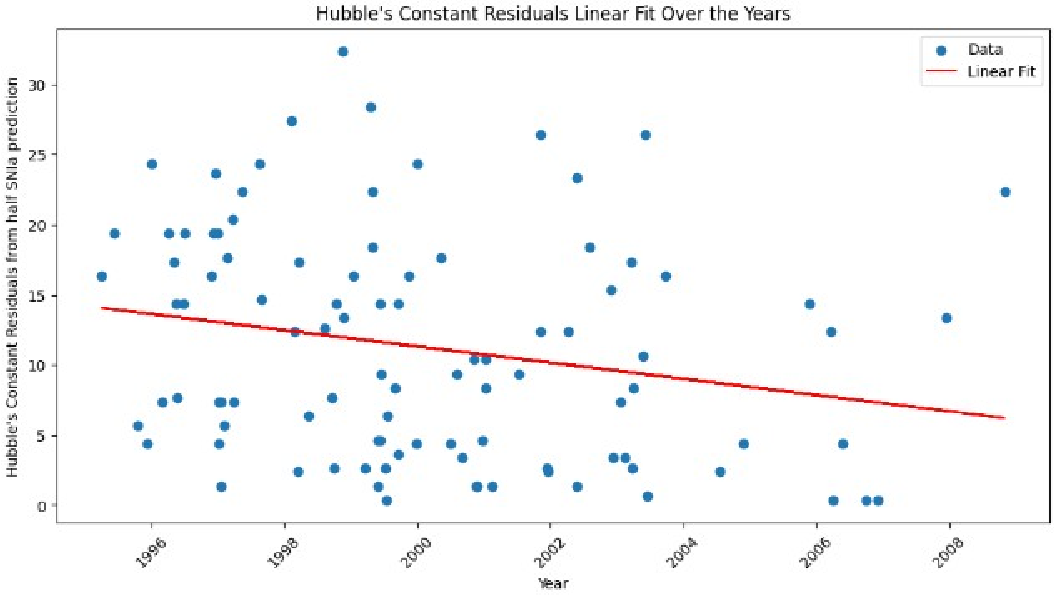


Some experimental types have notably exteme ranges, such as O, suggesting it is not a suitable method for extrapolating an ideal value of H. Some types only appear once in the larger data set, it is not possible to make meaingful evaluations on the suitability of those types other than comparing their one value.

It is possible that the quality of the data improves with time. Since we will extrapolate an ideal value of Hubble’s constant from the data set, we would only want to use the most recent academic data. Hence, it will be worth investigating the effect of time on the spread and accuracy of the data.

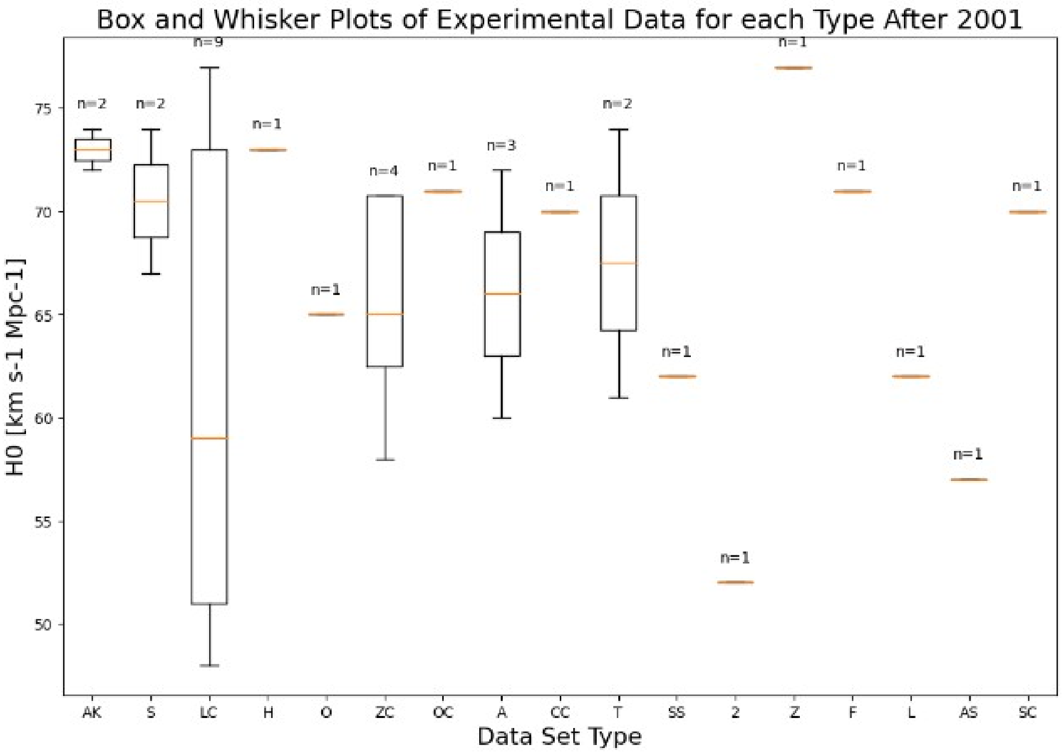


It looks like the precision and accuracy of measurements generally improves with time. This can be demonstrated by using a linear fit on a residuals plot of the literature Hubble 2 value and our calculated preliminary ideal value from Hubble 1.



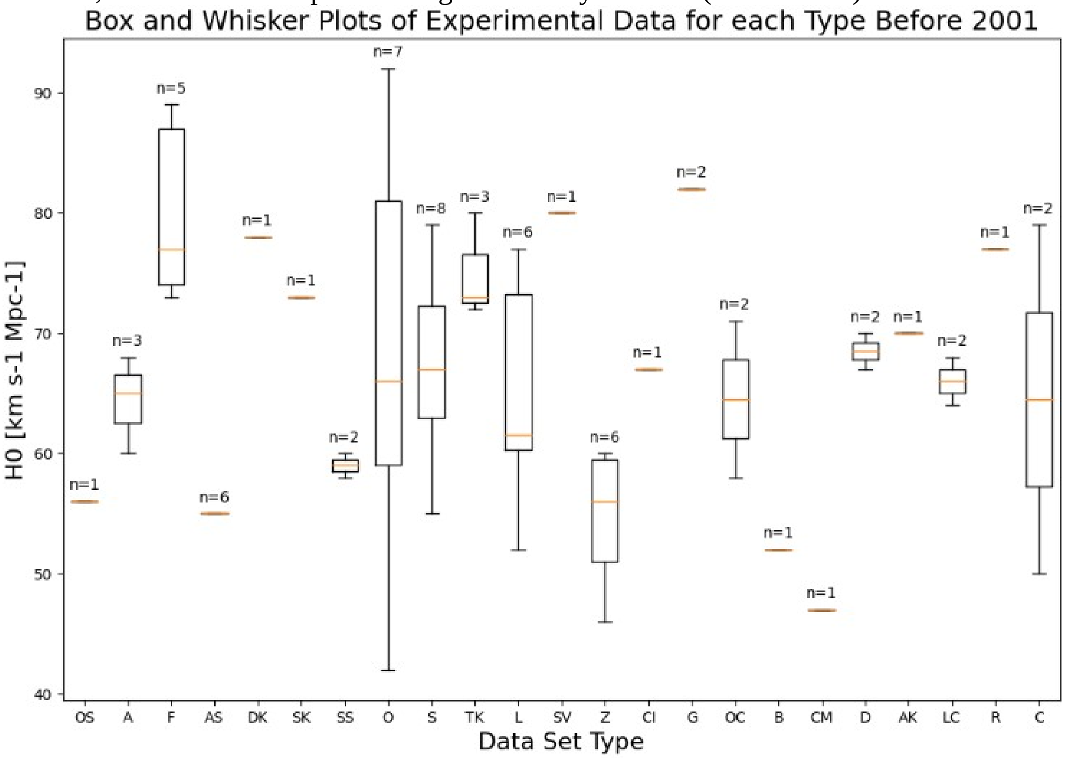
The difference between our ideal value and the literature value decreases with time. Despite this, the mean value for recent measurements of H0 is the same as the all time mean, suggesting that perhaps Hubble’s constant experimental data was well founded even in 1996.

It is possible to inspect the relationships between type and date by creating another multiple box plot, but using only modern data i.e. past 2001.



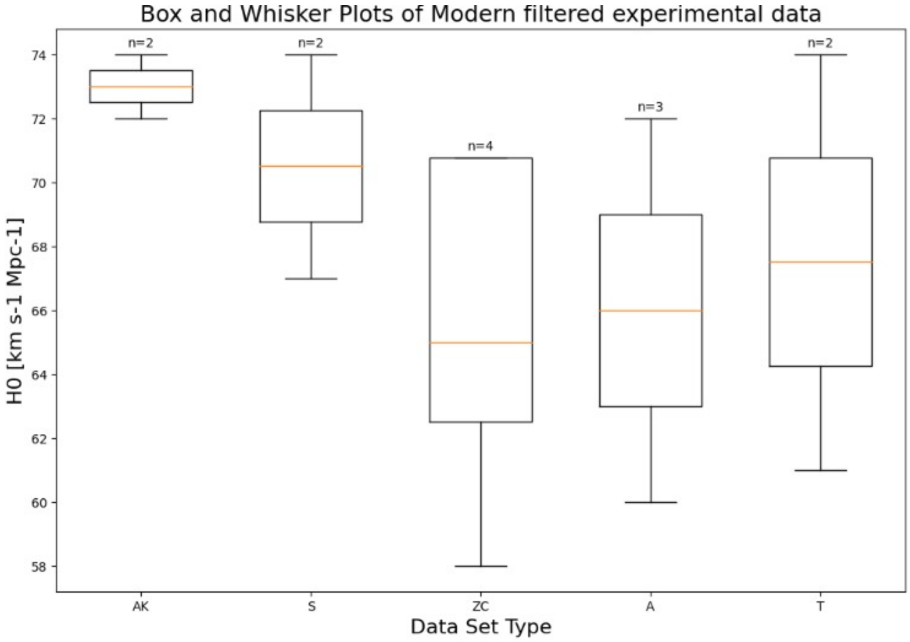
This data set has a mean H0 value of 65.82 km/s/Mpc with a standard deviation of 9.02 km/s/Mpc. Type LC measurements have a very large spread, suggesting LC experimental type is poor. The other modern experimental types present with a much reduced spread over the older data (shown below). For those experimental types where the sample size is 1 it is difficult to comment on the quality of the type.

For completeness, below is the box plot holding exclusively old data (before 2001).



This data has a mean H0 of 66.52 with a standard deviation of 12.27. Depsite the mean value itself being of a marginally higher agreement with our part a data, the spread of measurements has increased by a significant factor. Within this dataset we can see some experimental types are less suitable than others. For example, type ‘O’ measurements have very large degrees of spread. Type F has consistently poor accuracy, suggesting it holds an experimental error. Type S, L and C have a larger than ideal spread as well.

As a final effort to extract an ideal value of Hubble's constant from the experimental data, a filter was applied to modern data. The LC data presents with an abnormal spread compared to other data sets, suggesting it is a poor measurement type, so it was filtered. Experiment types with a sample size of 1, do not have a suitable vector to assess the quality of their data, since spread is meaningless, they were filtered as well.



The resulting dataset contains various experiment types with a consistent mean and spread, suggesting the experiment types are of similar quality.

The modern filtered data set has a mean value of H0 as 68.69 km/s/Mpc with a standard deviation of 7.42 km/s/Mpc. It is worth noting that this is remarkably close to the value from the European Space Agency's Planck mission, which produced "The most detailed map of cosmic microwave background to date" which was used to calculate a "Most likely" Hubble's constant of 67.8 km/s/Mpc. The value of 68.68 km/s/Mpc which was calculated from a strategical filtering of the filtered data set is henceforth the benchmark for Hubble's constant within this investigation and will be referred to as the ideal value.

Comparison to Part A Findings

The Tully-Fisher method has the most inaccurate value of H0 as 79.48km/s/Mpc with a standard deviation of 9.6km/s/Mpc. This method was thought to be inferior because of its lack of linear behaviour to support Hubble's relation. A linear regresion model yielded a regression coefficient of 0.778, numerically demonstrating the weak linear behaviour. The slightly reduced H0 accuracy to supports that the Tully-Fisher method is weaker. On the other hand, the model does admit to having a large spread so that the ideal value is still within 1 standard deviation, unlike for SNIa. The linear regression model yielded an additional way to find H0, as the gradient of the predicted data set. For Tully-Fisher this slope value was 77.9km/s/Mpc. This value has no associated spread making it less useful for analysis though.

The SNIa method overall has a slightly more accurate mean H0 value of 78.2km/s/Mpc with a lower standard deviation of 4km/s/Mpc. From part a this method was thought to be overall superior to Tully-Fisher because it obeyed a more visually linear relationship of Hubble's constant, and did so with a larger sample size, further demonstrating the validity of the relationship. In hindsight we can see that even ideal empirical data will have a large spread for Hubble's constant due to the multiple complications in taking measurements used to calculate the velocity and distance of astronomical objects.

Upon application of a linear regression model, the stronger linearity was demonstrated with a regression coefficient of 0.99. This suggested that the method would be suitable for finding a H0 value since it strongly obeyed Hubble's relation. The value of H0 given by the value of the linear regression gradient is 83.9km/s/Mpc which paradoxically deviates from the ideal value more than the Tully-Fisher linear regression gradient.

In part a, after seeing that Hubble’s constant appears to vary between two linear regions, the SNIa data set was split into two parts. Low distance (distance < 500 Mpc) and large distance (distance > 500Mpc).
The short distance part was found to have a mean H0 value of 71.8km/s/Mpc with a standard deviation of 1.2km/s/Mpc. This value is the most accurate so far to the ideal value, although the model has a very low spread which puts the ideal value within 2.6 standard deviations which is a significant disagreement. The value of H0 provided by the gradient of the linear regresssion model was 74.36km/s/Mpc.

The large distance data set was found to have a mean H0 value of 79.0km/s/Mpc with a standard deviation of 3.4km/s/Mpc. The fact that the mean diverages and spread increases for higher distances supports the previous understanding that large distance data have reduced precision and accuracy as a result of measurement issues due to cosmic dust and varability of the CMBR.

To conclude the comparison. The part Hubble 1 data only draws upon two different experimental methods, whereas the Hubble 2 data draws from multiple experiment types over many years. Even after filtering the data to only use the highest quality in finding an ideal value, this is still true. The ideal value from Hubble 2 is calculated from 5 different experimental sources in recent years, wheras the ideal Hubble 1 value is calculated from 1 experimental method. This suggests the stratified Hubble 2 data is a stronger source than Hubble 1.

Within the ideal experimental method used in Hubble 1, we identified a source of error due to the effect of cosmic dust. This was evidenced after applying a linear regression model. We found the velocity residuals appear to increase exponentially with distance after the training distance region is left. Cosmic dust is more likely to influence light, and also is more likely to have a stronger influence on light at greater distances between the object and the sensor. This explains the shape of the velocity residuals graph, whereby both the likelihood and amplitude of a deviation from the linear expectation is proportional to the distance.

The ideal value extrapolated from the Hubble 2 dataset is more accurate to the accepted literature value than that of Hubble 1, and the Hubble 1 data is subject to experimental error even within the most suitable data.