Error Analysis of *in situ* Sea Surface Temperature Data: Proj01

Background

Sea surface temperature (SST) is measured in this project by observing both *in situ* and satellite measurements. For each *in situ* dataset it is accompanied by a satellite measurement for the same longitude and latitude. Three datasets for *in situ* are considered: ship data (SH) operated by commercial, research, and governmental ships; drifting buoys or floats/ drifters (DB) that follow current paths in the ocean; mooring buoys (MB), which are anchored in a position. Each of these datasets share the same attributes i.e. columns as labels for the data entries. The attributes of interest are ISST, OSST, ICflag, OERR, year, LAT, and LON.

The ISST is the *in situ* measurement for the dataset. More on how each *in situ* is taken can found in the references [1]. The OSST is the satellite measurement taken of the heat of the ocean for the same location. The OSST is a remote sensing technology and this data is appended to each *in situ* dataset for comparison. The ICflag is the *in situ*'s quality of the measurement. The ICflag is binary. 0 is marked for bad measurement while 1 is noted for good quality measurement. The OERR is the error estimate associated with the satellite measurement. The OERR estimate is quantitative (ranges from 0 to 1). The year is the annual cycle in which the measurement was taken. LAT represents latitude. LON represents longitude.

Exploratory Data Analysis (EDA): data wrangling, cleaning, and surveying.

For open-ended projects it is common practice to investigate and shape the data to verify it is in the form as expected. For the ISST datasets we check for uniformity for the attributes of interests to be able to compare behavior between groups. This step of pre-processing done with diligence can generate research questions by noticing what is the same and what is different between the data sets. The tools and packages for EDA can be found in the online code reference [2]

Shipping Data (SH)

Surveying the shipping data (SH) shows its path and what seems to be docking stations. From fig. 1 we observe that the range of data in question are from 1991 to 2006.

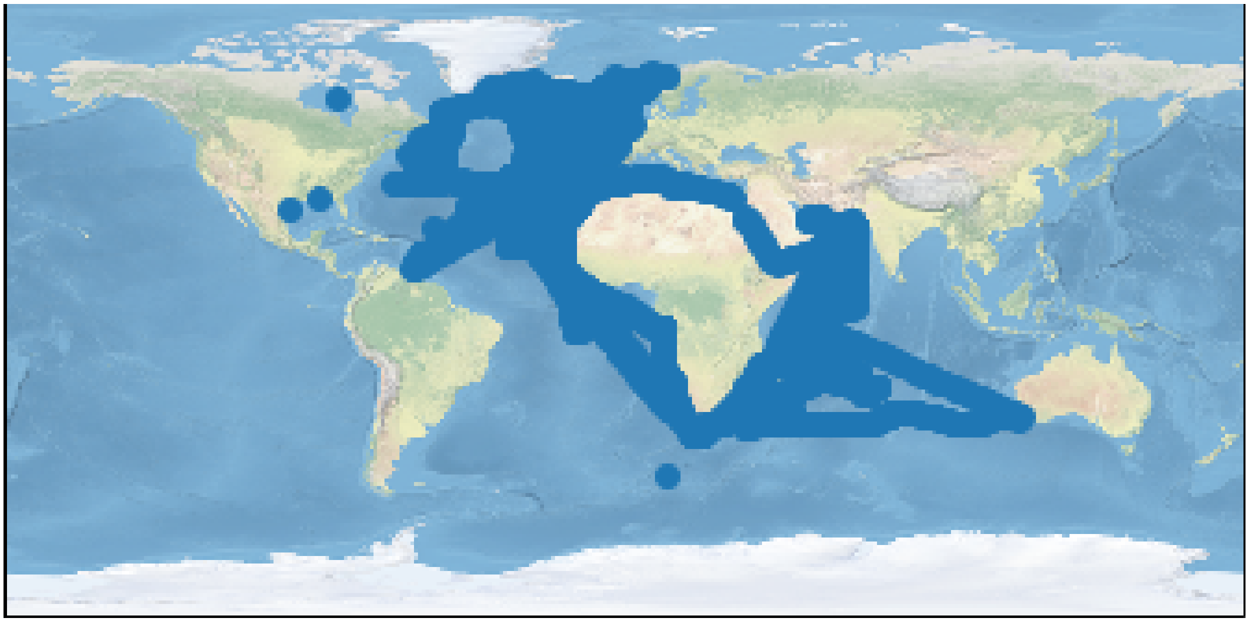

Fig. 1: Shipping Data path from 1991 to 2006

Drifting Buoy (DB)

For the drifting buoy (DB) our dataset ranges from 1995 to 2010. We see that the DB in question is near the sub-arctic and follows the current along that path without touching significant land mass between Australia and the tip of Antarctica.

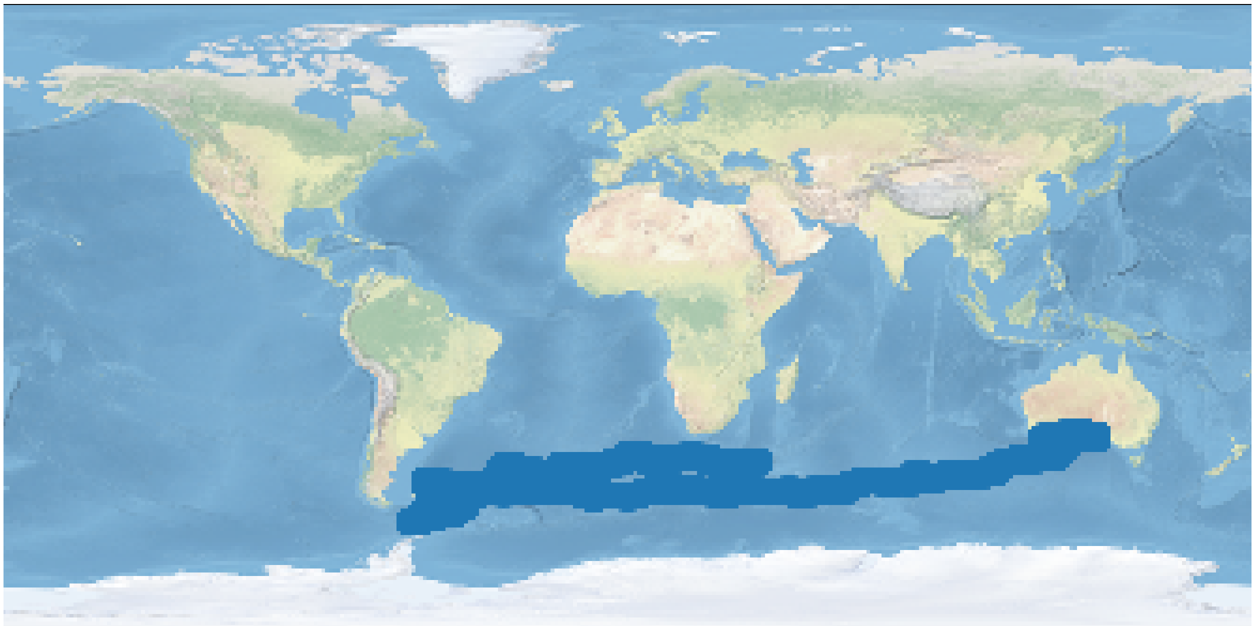

Fig. 2: Drifting Buoy path from 1995 to 2010

Mooring Buoy (MB)

For the mooring buoy (MB), a first glance suggests many locations for the range of years between 1991 and 1996. Fig 3 shows all of the locations superimposed on one another for all possible points.

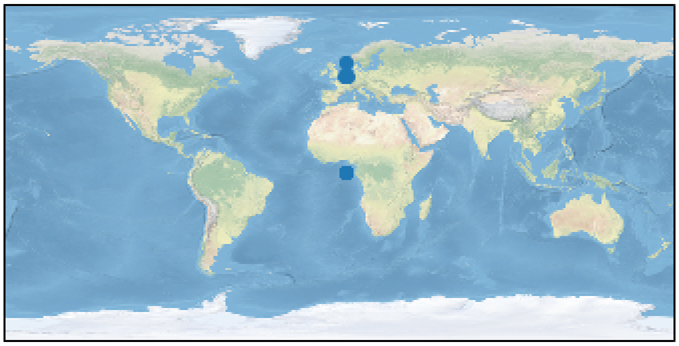

Fig. 3: Mooring Buoy locations from 1991 to 2006

The following table shows the count of all the locations suggested by the data for docking of MB sorted by year.

Year	No. of locations
1991	2
1992	2
1993	5
1994	3
1995	2
1996	1

Fig. 4: Table of suggested no. of locations for mooring buoy by year

Analysis shows that there is one docking station of significance for all of the years except 1993. For 1993, results show that there are two objective locations labeled 1993A and 1993B while other locations can be regarded as noise. Kaplan explains the concept of “wiggle” for the mooring buoys [1]. The techniques outlined in this project combines data science and statistical analysis to objectively handle “wiggle”, filter out spurious results, and concludes true location of mooring buoys regardless of label with 95% confidence.

Analysis

Step 1: Graphical Filtration

Using techniques in data science, a preliminary filter is done graphically to generate an assumption on the true number of MB exist by year. Fig 5 shows a scatter plot on the globe for location of 1993 MB by changing the gradation to visually count the number of independent mooring buoys for a hypothesis.

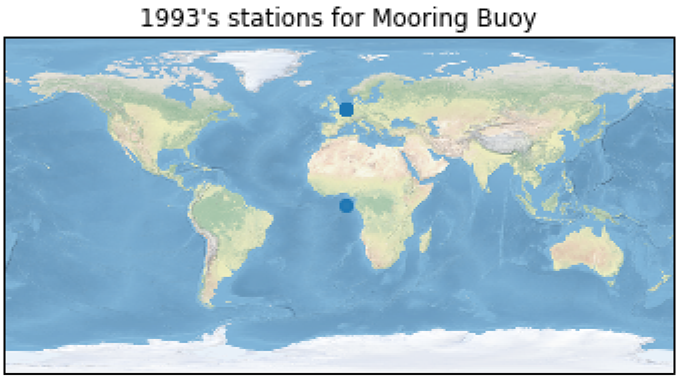

Fig. 5: 1993 hypothesis of mooring buoys

Applying an alpha = 0.1 shows that the year 1993 is the only year that has multiple stationed buoys. By inspection this year has two different buoys for this year while the rest has only one. The plots for the other years can be considered as "wiggle" or error in the data as there are not a significant number of points to display a mark.

Step 2: Generate hypothesis

By inspection there are two 1993s: 1993A (51.9°, 3.2°) and 1993B (0°, 3.2°). We can run a null hypothesis t-test on this assumption by comparing the difference between their means. Noting that all of the clean data agrees with a 3.2º line of longitude. We will take any latitude measurement above 25º as belonging to 1993A and measurements below 25º belonging to 1993B. We can run this test for both OSST and ISST. We will assume a normal distribution of data and for each 1993s we will compare the mean for the entire data set to the mean of the assumed location. Our confidence interval will be 95%.

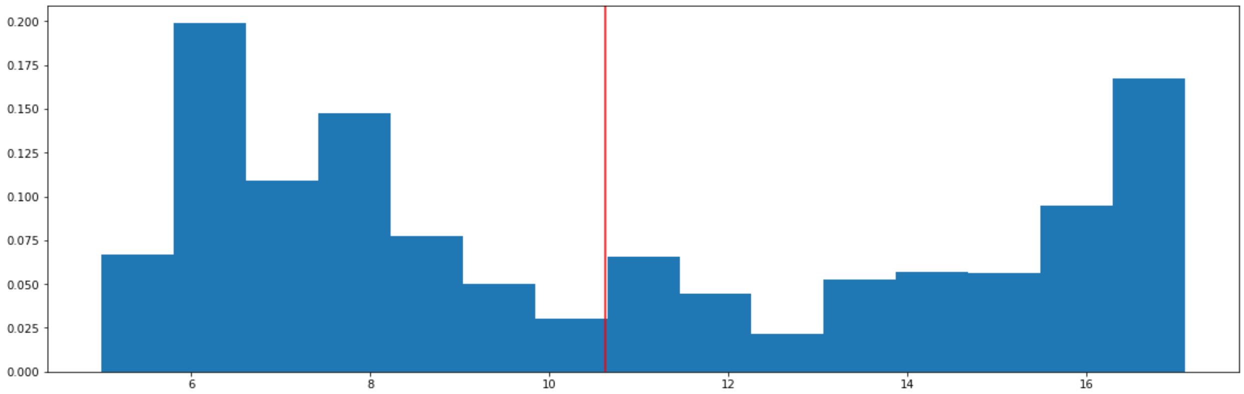

Fig. 6: 1993A distribution for ISST

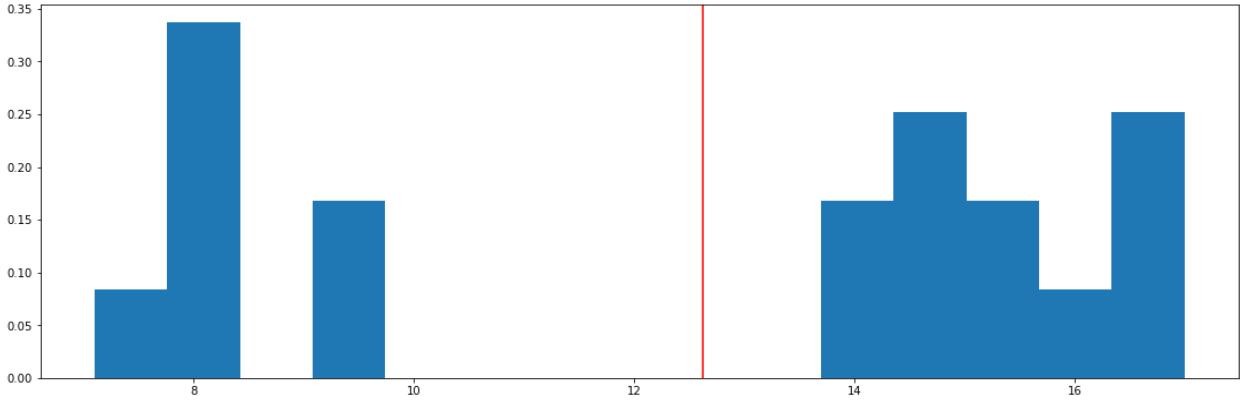

Fig. 7: 1993B distribution for ISST

The mean for 1993A is 10.627 (where all measurements are in ºcelsius) for 1607 points above 25º latitude. The mean for 1993B is 12.628 with 18 points below 25º latitude. We run a t-test for hypothesis testing and fail to reject the null hypothesis with a p-value of 0.03. As expected ISST 1993A and ISST 1993B are indeed different. We did a hypothesis difference test with a 95% confidence interval. Note that for 1993B there are few points in our count so the t-test was appropriate.

Step 3: Validate results

We run the same test, but this time using the OSST's measurements. 1993A has a mean of 11.087 and 1993B has a mean of 26.67. Again we fail to reject the null hypothesis with a p-values ~ 1.6 x 10^-51.

Again, for similar reasons we are able to establish that there are two different MOORING BUOYS for 1993 under OSST measuring instrument. This leads to further reliable results as ISST and OSST has different thresholds for measuring quality of measurement. We are confident that 1993A and 1993B are different locations. Now we can test for each location whether our filter criteria for longitude and latitude is representative of the pool of data or an aggregate of independent nearby measurements. We will test whether coordinate (51.9°, 3.2°) is representative of 1993A and whether coordinate (0°, 3.2°) is representative of 1993B. We consider ISST measurements.

Step 4: Refine, sampling distribution

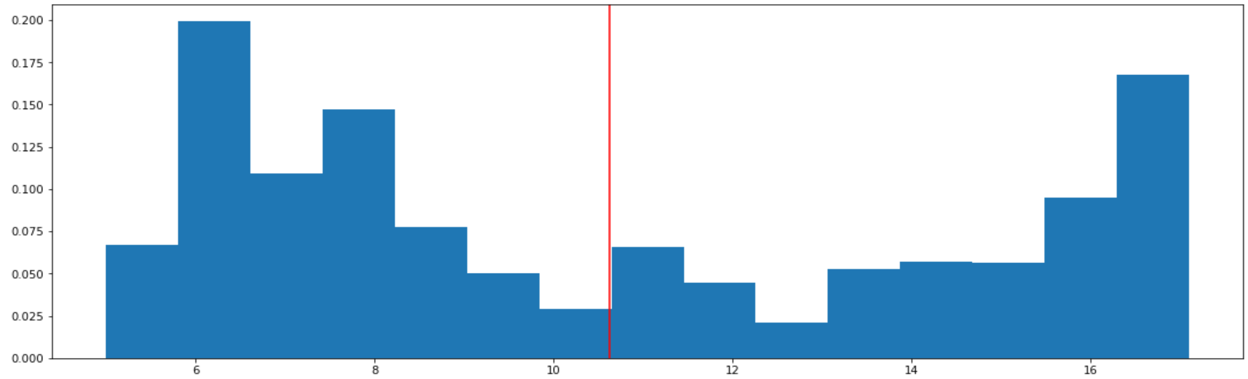

Fig. 8: 1993A sampling distribution for ISST

The mean for the sample shown in fig. 8 is 10.99 with 1604 points considered out of the pool of 1607. Already it is clear that our filter and specifically the coordinate for 1993A is accurate. There aren't enough points to boot-strap a distribution for an alternate hypothesis. Coordinates that don't align with (51.9°, 3.2°) are spurious. We obtain the same results when running the script for OSST because the coordination of these points are chosen to match that of ISST– the data was cleaned to filter out rows that had missing values for either of these. If there were a considerable amount of data that were dropped we could have done a statistical test on these as well to test the validity of our cleaning method. The amount of data that were dropped is 1 row and that was from ISST: we didn't consider the OSST from that row. We dropped this row for uniformity of code and maintain reliability of representing the data provided.¶

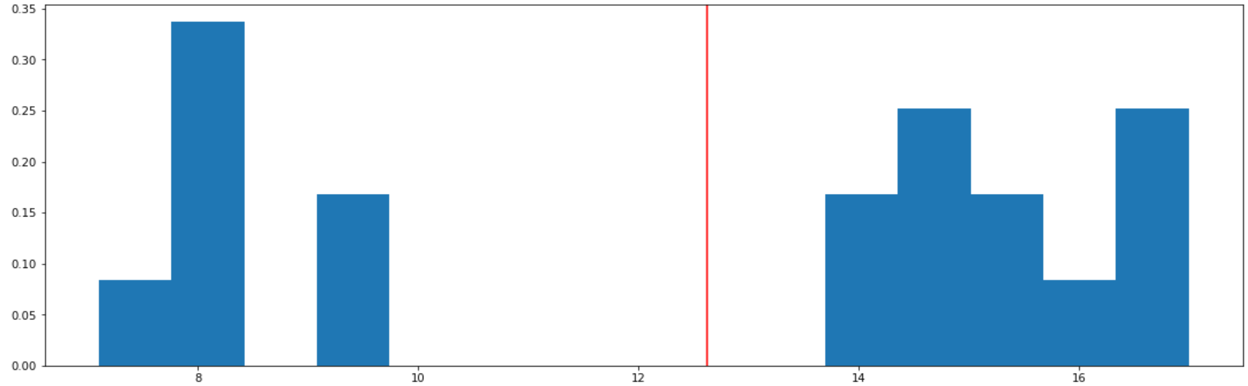

Fig. 9: 1993B sampling distribution for ISST

For 1993B, all the points are represented by coordinate (0º,3.2º). This case is trivial. We have 100% reliability.

Conclusion

We separated the data by years and focused on tracking our Mooring Buoys. Originally we found that multiple years had multiple locations. Using a graphical filter we filtered out spurious locations for a great many years. We were left with two locations of the mooring buoy for year 1993 from an original set of 5. From this we were able to show that these two locations were statistically different based on their means when we ran a null hypothesis. Thirdly, we validated these results by using an addition degree of freedom: namely, we ran the test twice and got the same results for ISST and OSST.

Ultimately, we have shown that data science techniques for data cleaning can follow statistical analysis for judgement. Using a graphical plot we can filter out first layer noise. As a rule of thumb, of the number of points filtered out graphically are so few that boot-strapping a distribution would be inappropriate, then the data filtered out can be regarded as noise. As a second layer, run statistical analysis to verify whether your points are different. Since we only had two points to consider for 1993, we ran a simple hypothesis test. If there were more than two points we would have ran an ANOVA test. A pairwise null hypothesis test can be done to identify which points are different, if an ANOVA yields results that points are different. Validating results can be done by running analysis again for the number of attributes you have in your data set. In this case we had two: ISST and OSST. For similar testing technique of our locations, we can run null hypothesis and ANOVA testing on our attributes.

A general technique has been mapped out to analyze locations of mooring buoys despite the years that they are labeled. This process is objective and therefore can be developed into a script for meteorological scientist to make unbiased parsing of their data without having to rely on documentation from various communities.

References

[1] Kaplan, Alexey, Error Analysis of in situ Sea Surface Temperature Data: Introduction into Students' Projects, 2019, drive.google.com/drive/u/0/folders/13DDg7q5cvoY56ryGkC6Nf9tqcUU_HQHy

[2] Adomako, Ronald, github.com/adomakor412/Applied_Statistics/blob/master/Final/Final_Project_1_data_set.ipynb [accessed 12/17/19]

[3] International Comprehensive Ocean-Atmosphere Data Set, https://icoads.noaa.gov/, https://icoads.noaa.gov/r2.5.html [accessed 12/17/19]

[4] International Research Institute, Lamont-Doherty Earth Observatory http://granger.ldeo.columbia.edu/ [accessed 12/17/19]