Report 2

Dataset: ‘El Nino’

Context: El Nino as a subject is a cyclic fluctuation of the temperature of the sea surface and air temperature covering the entire equatorial Pacific Ocean. This phenomenon has a direct influence on the weather across the region and other parts of the globe. The correspondent dataset was sourced from the Tropical Atmosphere Ocean (TAO) which uses a network of floating buoys over the equatorial Pacific. The buoys systematically collect data related to oceanographic and surface meteorological variables aiming to understand and guess the cyclic variations in climate associated with El Nino. By analyzing this dataset, the concerned can gain valuable insights into the phenomenon events and their broader climatic implications.

Data Description: the variables measured by the buoys are, ate, latitude, longitude, zonal winds (west<0, east>0), meridional winds (south<0, north>0), humidity, air temperature, sea surface temperature, and subsurface temperatures. The longitude and latitude refer to the position of each buoy, also the dates represent the periodic sequences of measurements since 1980. The data contains missing values due to practical issues.

Data validation:

The following table represents a checklist that profiles the dataset including cleaning if needed to prepare the data qualitatively.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Checkpoints | Validity (V/NV) | Number of issues | Observation | Handling | Result |
| Readability | NV | 2 | Data provided in table format + no col names | Conversion into CSV format + assigning the column names from (tao-all2.col) | El-nino.csv |
| Data type conversion | NV | 1 | Date | As.Date(date) | Date type |
| Missing data | NV | 5 | *5 variables that include N/A* | Identification of missing values to consider later+ replacement by(.) | Identified missing data. |
| Incoherent data | NV | 2 | Date + longitude  format | Date format conversion "%y%m%d" +  tao$longitude%%360 | Coherent data |
| Potential Outliers | V | - | - | - | - |
| Duplicates | V | 0 | - | - | - |

Data exploration

Since the data has an important amount of missing data, the following step has been taken to measure the unavailable data.

* Missing values quantification

|  |  |  |
| --- | --- | --- |
| Variable | Number of N/A | Percentage\_Missing |
| zon.winds | 25163 | 14.13 % |
| mer.winds | 25162 | 14.129 % |
| humidity | 65761 | 36.92 % |
| air temp | 18237 | 10.24 % |
| s.s.temp | 17007 | 9.55 % |

More than 36% of humidity data is missing, and you can see that the number of missing values can be significant in overall data. So, this has to be considered during the analysis.

* Summary (tao)

A close-up of a list

Description automatically generated

Based on the data summary, observe that the variable ‘year’ extends from 1980 to 1998 and covers all months of each year as shown in min, and max of variable ‘month’, this suggests potential seasonal trends associated with meteorological variables. Nonetheless, the ‘humidity’ variable has a close median to the mean.

* Correlation inspection

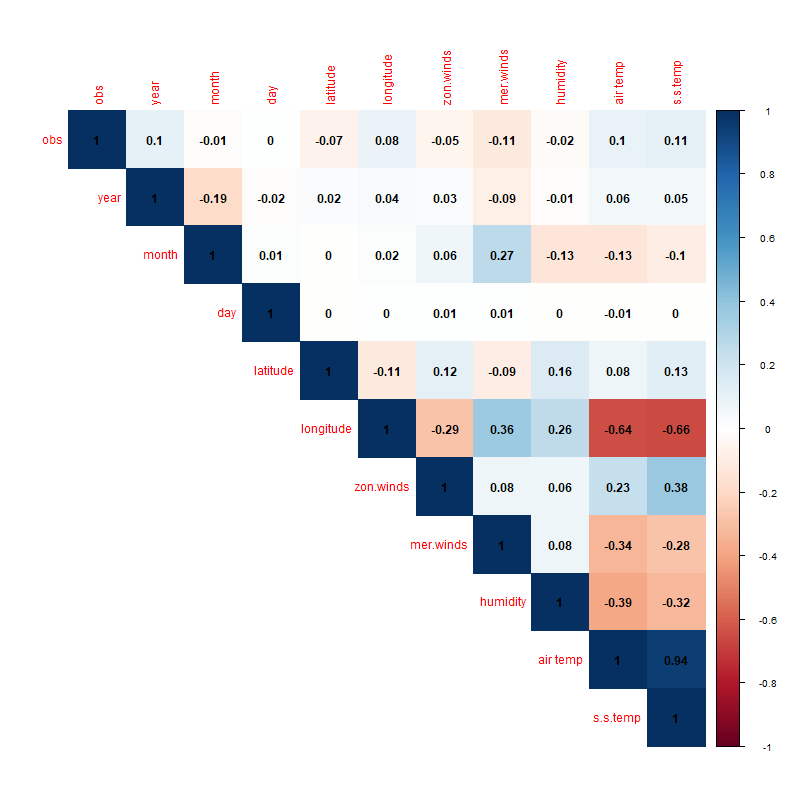


Figure . Correlation heatmap matrix

Upon inspecting the correlation heatmap matrix, there is a strong negative correlation between the variable ‘longitude’ versus ‘air temperature’ as well as between ‘sea surface temperature’ versus longitude, in addition to a stronger positive correlation between the ‘air temperature’ and ‘sea surface temperature’. These observations indicate a significant covariant relationship between longitude and both temperatures.

* Relationship analysis

In order to extract insight into the variables that show correlation in the heatmap above, the following scatter plot illustrates the relationship between ‘s.s temp’ and ‘longitude’ .

A graph of a graph showing a temperature

Description automatically generated with medium confidence

Figure . S.S temp and longitude relationship

Visually, the plot exhibits a clear variation in terms of sea surface temperature depending on the longitude value where as long as the position moves towards the east the temperature tends to be cooler with an S.S temperature interval of [18, 31] °C. Conversely, moving into the west tends to have a warmer temperature where the S. S temperature varies between 24 to 31 °C.

Concerning the air temperature measurement, the next plot visualizes the variation of that temperature associated with longitude.

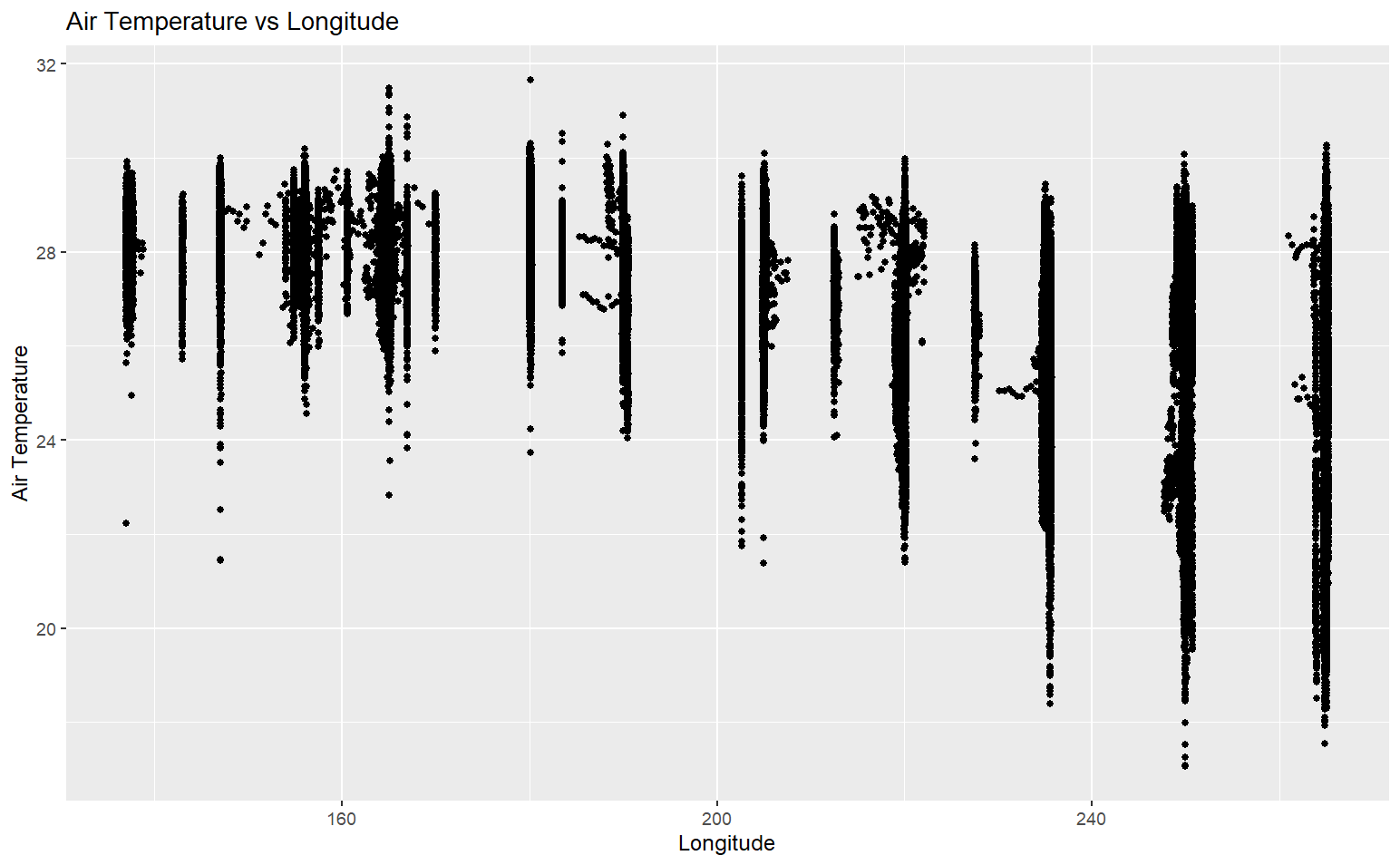


Figure . Air temperature versus Longitude relationship

Similar features are observed in terms of air temperature where it spans larger when longitude positions towards the east which contextually indicates cooler air temperature and warmer air temperature around the western direction.

Accordingly, and in order to emphasize the previous discussed features, the following scatter plot visualizes the ‘air temperature’ and ‘S.S temperature’ variables with data points colored based on their longitude value.

A graph showing a black and grey line

Description automatically generated with medium confidence

Figure . S.S temp and Air temp relationship

The western longitudes are marked by the gradual lightest blue color indicating the higher temperatures of both measures air and sea surface compared to the eastern longitudes which have cooler temperatures colored in gradual black, This pattern underscores the strong positive correlation observed during the correlation inspection.

* Time series

Another aspect of analysis that explores the relationship between the subsurface temperature and air temperature variations is their fluctuation over a periodic cycle which is a subject of analysis for this data set to understand the impact of the environmental variables on the El Nino effect. The subsequent plots present time series representations for zonal winds, meridional winds, humidity, air temperature, and subsurface temperature. To delve into this relationship while mitigating the impact of missing data mentioned earlier, a subset of the dataset was employed. Specifically, the year 1994 is chosen as a case study period in the following figure.

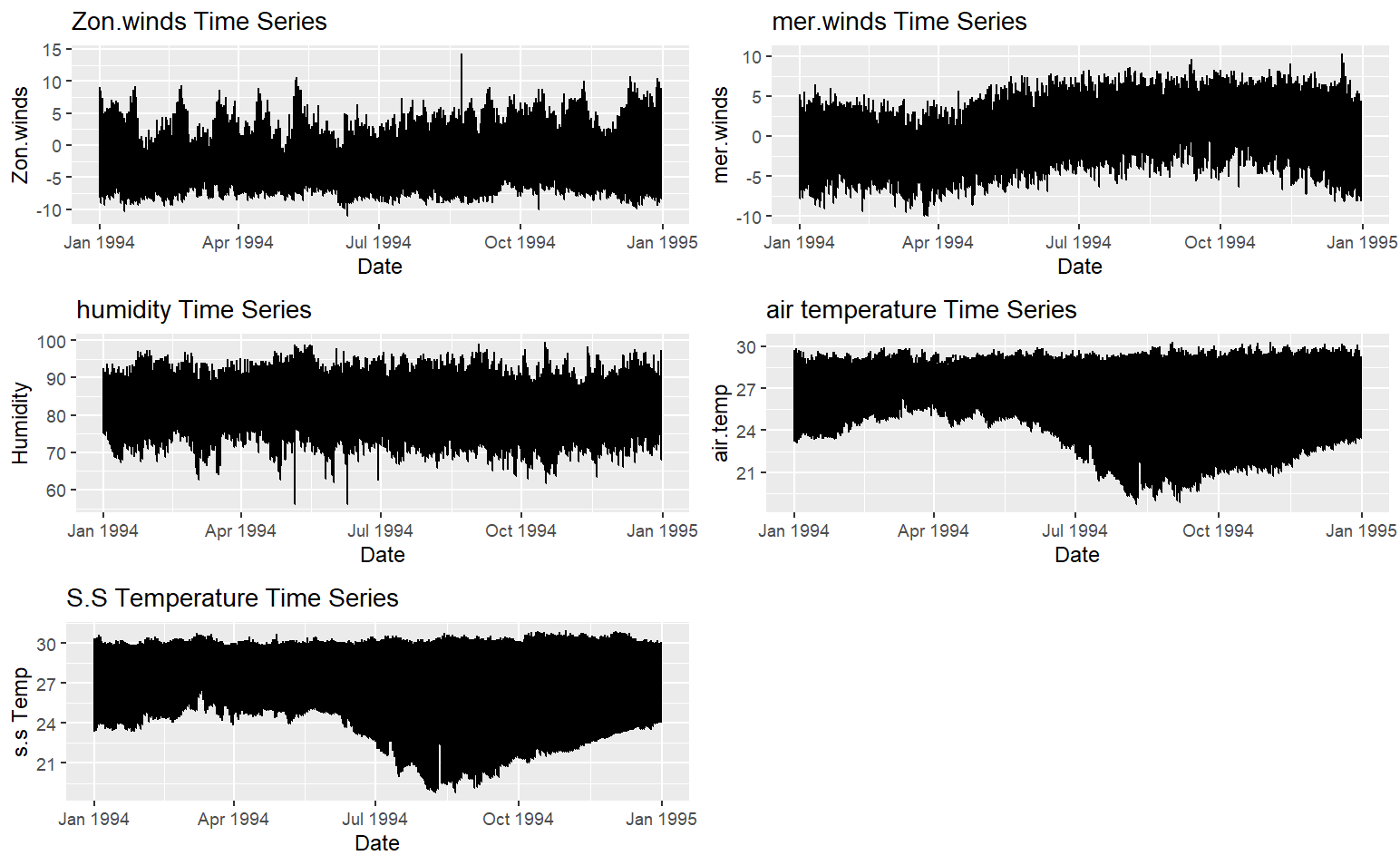
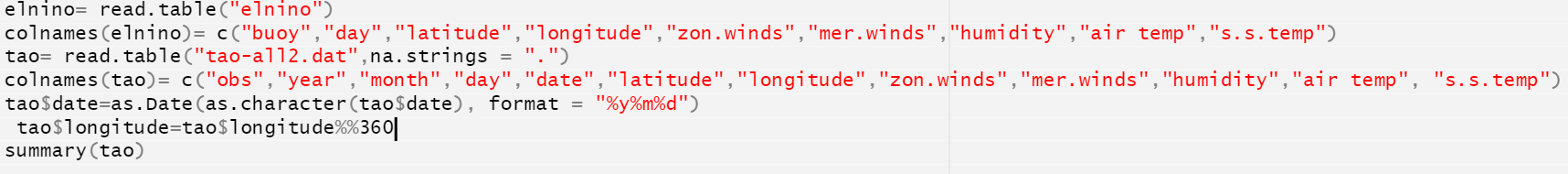
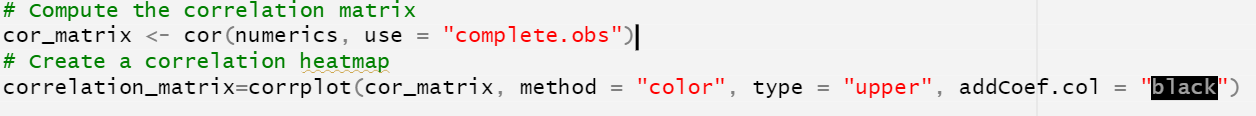


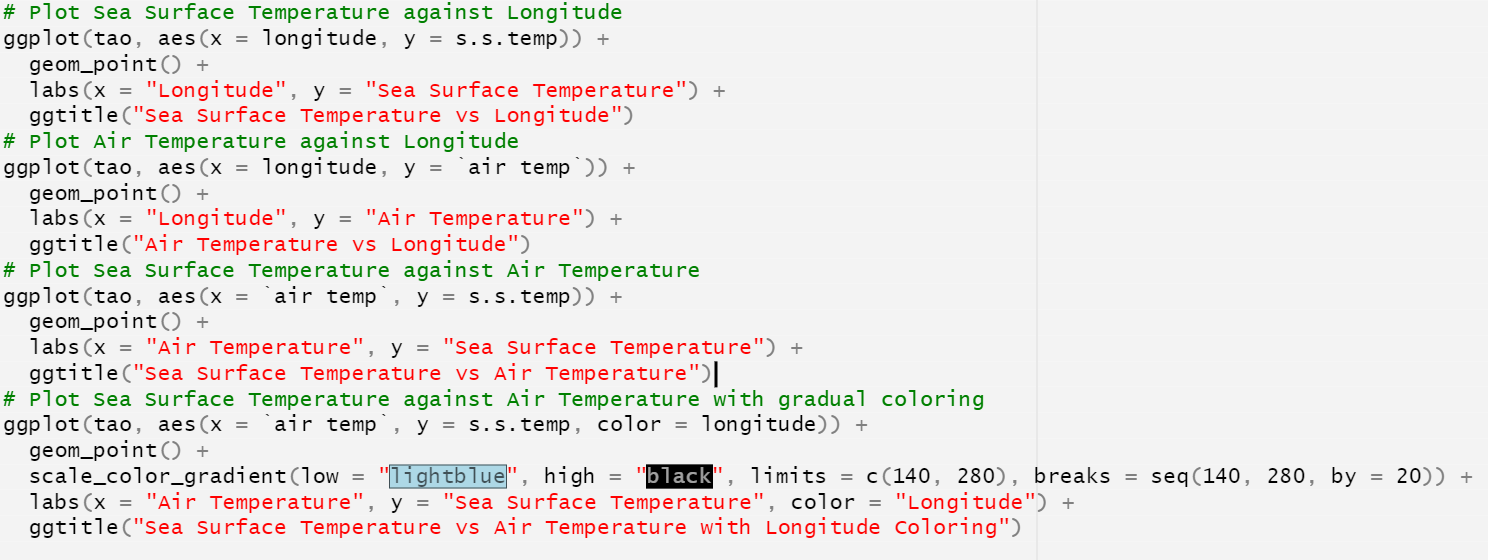
Figure . Time Series plot year (1994)

Key trends derived from the time series visualization reveal a notable similarity in the behavior of air temperature and subsurface temperature over time. Both temperatures exhibit a decreasing trend from July to October, followed by a gradual increase towards the end of the year. This pattern underscores the correlated relationship between these variables highlighted earlier. On the other hand, the meridional winds present a distinct pattern, with more pronounced fluctuations in the southern region during the first quarter of the year (January to April). Subsequently, there is a continuous increase until October, followed by a subsequent decrease.

Code:







A screenshot of a computer code

Description automatically generated