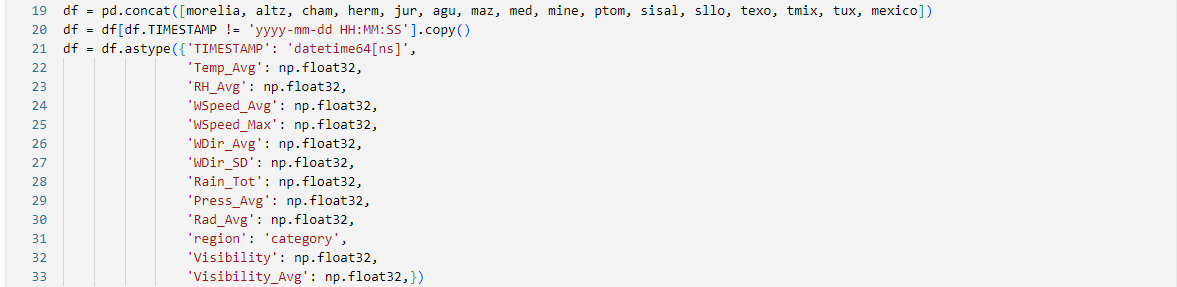
# Conversion initial data

Initial data contains almost 1500 csv files. With information about pollution and meteorology data. Initial data contains 12 columns and we have added a new column – region (Region of our station) and united all tables in one. Also, we set dtypes of each column.

|  |  |  |
| --- | --- | --- |
| Name | Dimension | dtype |
| TIMESTAMP | yyyy-mm-dd HH:MM:SS | Datetime64[ns] |
| Temp\_Avg | C | float32 |
| RH\_Avg | % | float32 |
| WSpeed\_Avg | m/s | float32 |
| WSpeed\_Max | m/s | float32 |
| WDir\_Avg | deg | float32 |
| WDir\_SD | deg | float32 |
| Rain\_Tot | mm | float32 |
| Press\_Avg | hPa | float32 |
| Rad\_Avg | W/m^2 | float32 |
| Visibility | M | float32 |
| Visibility\_Avg | M | float32 |
| Region | Region Name | category |

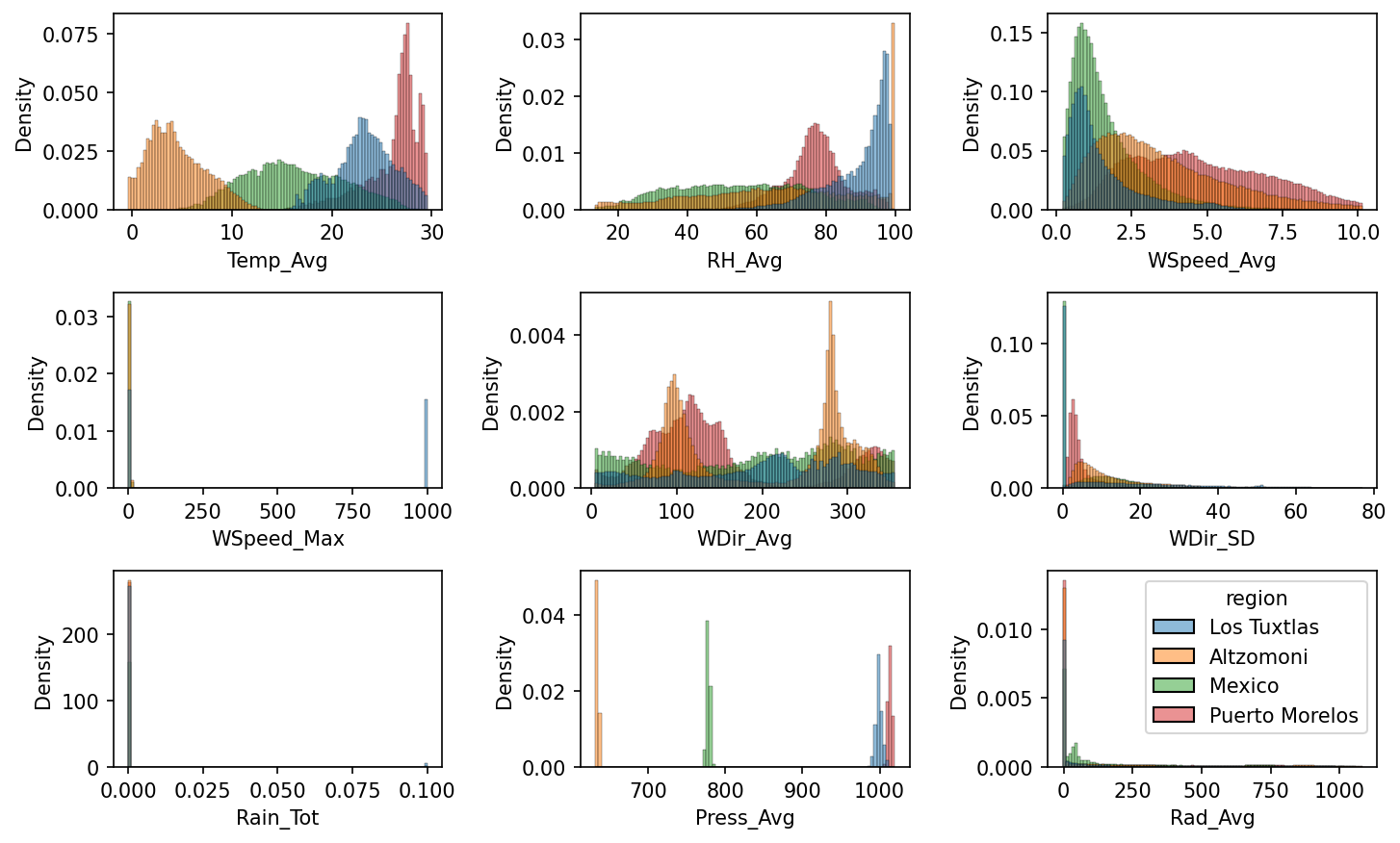
As a result we decreased RAM from 5.8 to 2.9GB and disk space from 10GB to 1.4GB.





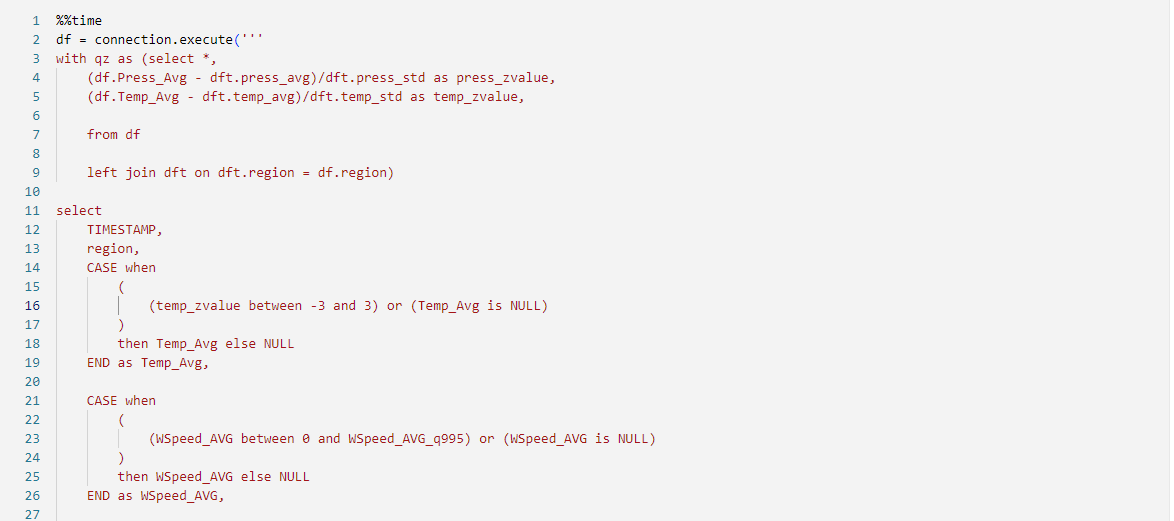
# Cleaning dataset

Distribution of each parameters should be cleaned in different regions separable. Based on distributions in different regions we decided to make this scheme of cleaning:



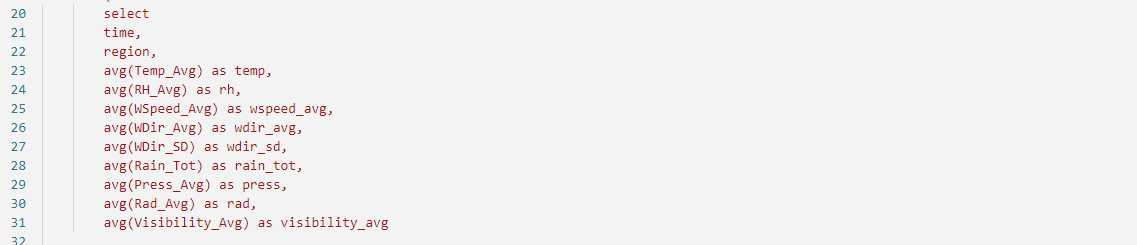
|  |  |  |
| --- | --- | --- |
| Name | Cleaning way | Why? |
| TIMESTAMP | - | - |
| Temp\_Avg | -3 <= Z – value <= 3 | Distribution looks like normal |
| RH\_Avg | 1 <= x <= 99 | There are outliers = 0 and 100 |
| WSpeed\_Avg | 0 <= x <= 0.995 quantile of x | Physically value and high outliers |
| WSpeed\_Max | 0 <= x <= 0.995 quantile of x | Physically value and high outliers |
| WDir\_Avg | 0.001 <= x <= 359.999 | There are outliers = 0 and 360 degree |
| WDir\_SD | 0.001 <= x <= 359.999 | There are outliers = 0 and 360 degree |
| Rain\_Tot | 0 <= x | Physically value |
| Press\_Avg | -3 <= Z – value <= 3 | Distribution is normal |
| Rad\_Avg | 0.001 <= x | There are a lot of distribution at x = 0 |
| Visibility |  |  |
| Visibility\_Avg | 1000 <= x <= 19500 | Outliers based on distribution |
| Region |  |  |

Cleaning was made using postgresql (duckdb)

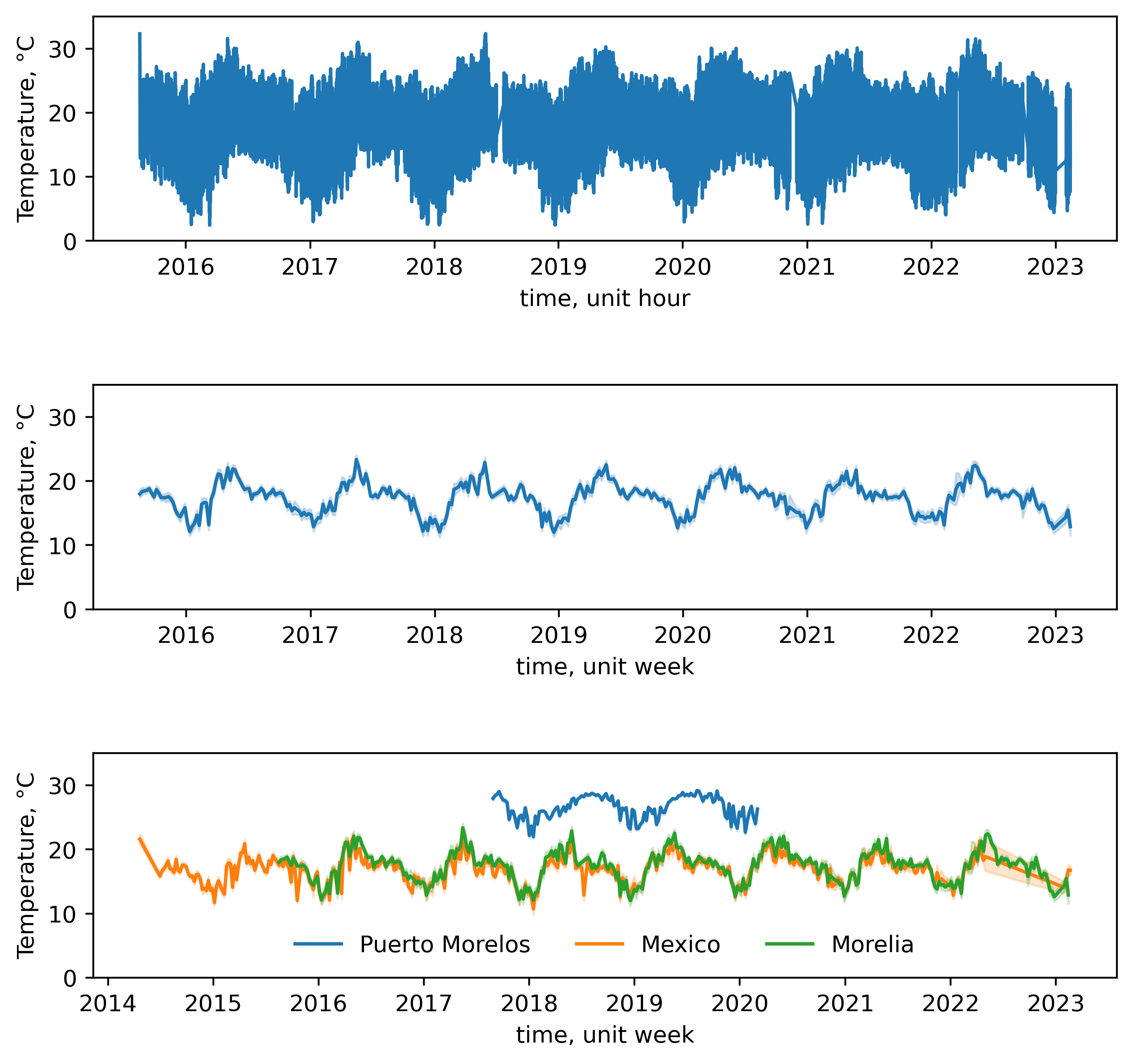


# Data analyses

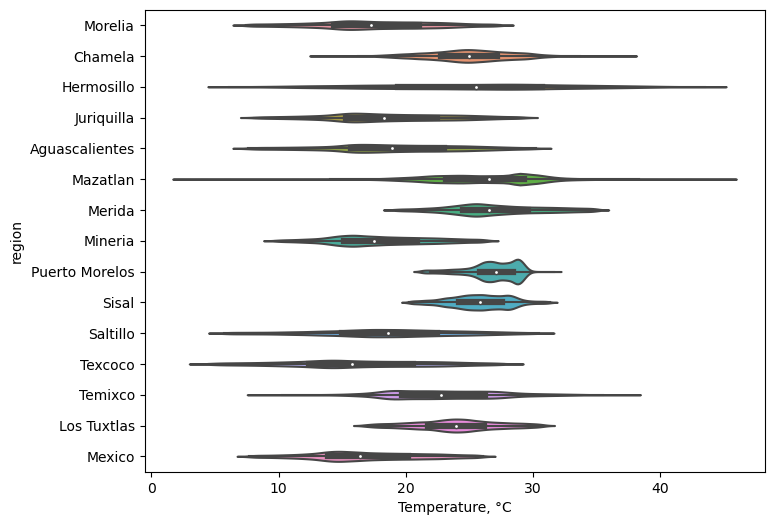
For faster analysis, we choose average values of parameters for hour.



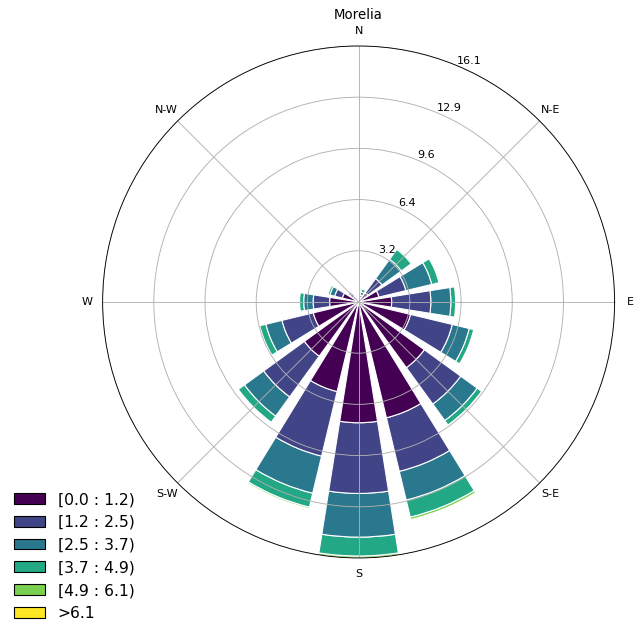
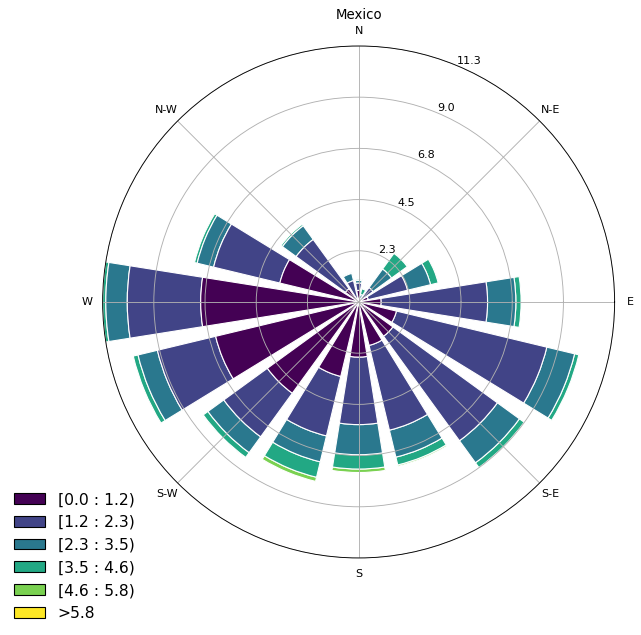
Based on this we can show time series plots of different fields for example temperature with averaging on different time steps. Also, we can compare different regions.



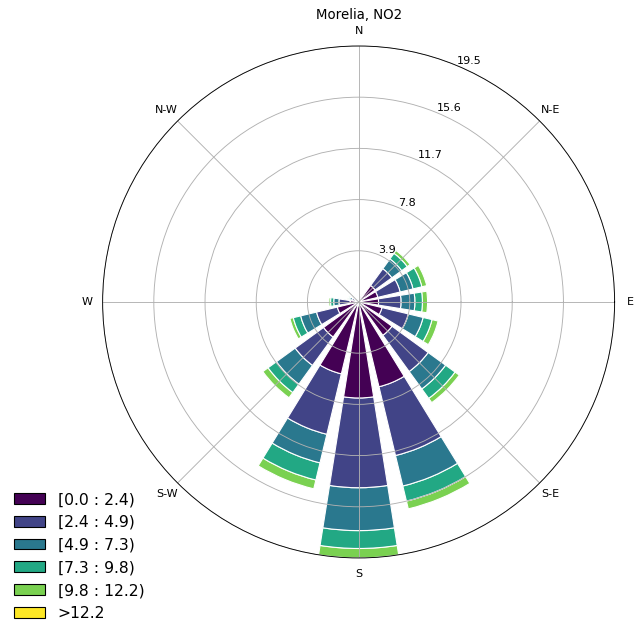
From this data, we can get distribution of different values and see how they change in different regions.



From data we can get wind Rose…

Wind rose with pollution



We can get more complex analysis of seasonality …



