Spatio-Temporal Analysis of America temperature in 2015

Zhu Lin

Getting startted

The film "the wandering earth", which has attracted a lot of attention recently, sets up an extremely harsh climate which is close to collapse after a great change of climate. Such a catactrophic scenario is not just a figment of the bad weather that is common in many places. Climate change and potential global warming have become the significant issue of the day, monitoring temperaturesacross the region, therefore, is becoming increasingly important in response to weather-related disasters.

The data files US_weather_2015 and weather_stations correspondingly contain temperature of the United States at different stations and and their locations (marked by longitude, latitude and elevation) from January 1 to December 31, 2015, a period of 365 days. Given spatial statistics models generally only take longitude and latitude into consideration, we substract the variable elevation (elev) and replace the temperature with the mean temperature(temp_mean) in the same time at the same location (including stn, lon and lat). Furthermore, to facilitate analysis, we should merge two datasets according to the same variable station number (stn). Top of all, we assume that temporal resolution is one day, and hence we are expected to divide the whole data into 365 sub dataframes including the information of each day, and denote them by "day1", "day2" and so on.

Exploratory data analysis

Data description

The following packages are used in this project.

```r
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':
##
##     filter, lag

## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```r
library(ggplot2)
library(maps)
```

Read two data files US_weather_2015 and weather_stations

```r
weather = read.csv("US_weather_2015.csv")
loc = read.csv("weather_stations.csv")
```

Dataframe weather has five variables: stn, year, month, day and temp.

```r
head(weather)
```

```
##      stn year month day temp
## 1 916520 2015     1   1 83.0
## 2 916700 2015     1   1 86.9
## 3 788460 2015     1   1 81.8
## 4 916600 2015     1   1 84.3
## 5 965650 2015     1   1 79.0
## 6 406370 2015     1   1 47.6
```

Dataframe loc has seven variables: stn, name, country, state, lat, lon and elev.

```r
head(loc)
```

```
##      stn                          name country state     lat     lon elev
## 1 423630 MISSISSIPPI CANYON OIL PLATFO      US    LA  28.160 -89.220   37
## 2 619760   SERGE-FROLOW (ILE TROMELIN)      US       -15.883  54.517   13
## 3 621010                   MOORED BUOY      US        50.600  -2.933 -999
## 4 621110                   MOORED BUOY      US        58.900  -0.200 -999
## 5 621130                   MOORED BUOY      US        58.400   0.300 -999
## 6 621140                   MOORED BUOY      US        58.700   1.100 -999
```

Merge weather and loc by common variables stn, lat and lon and remove duplicate rows based on all variables

```r
## merge weather and location(loc)
library(dplyr)
loc = select(loc,-c(name,country,state,elev))
data = left_join(weather, loc, by = "stn")
data = filter(data, is.na(lat) == F)
## Remove Duplicate Rows based on all variables
data = distinct(data)
```

Add some variables, do data cleaning and replace "temperature" with "temperature_mean".

```r
## add a variable "day" to identify date
data = mutate(data, num = as.numeric(ISOdate(2005,month,day)-ISOdate(2005,1,1),
                                   units="days")+1)
## add a variable "atr" to distinguish their attributes
data = mutate(data, attr = (stn/10000+lon/10+lat+num)*10000)
temp_mean = summarise(group_by(data, attr), temp_mean = mean(temp))
data_clean = left_join(data, temp_mean, by = "attr")
data_clean = select(data_clean,-c(temp))
data_clean = distinct(data_clean)
```

To familiarize ourselves with the geography of the dataset, we will initially ignore the temporal component of the dataset and examine the spatial distribution of temperatures on a single day. Take day 1 as an example, therefore, we extract data of January 1, 2015 and add variables lon_rank and lat_rank.

```r
## Extract day 1's data and sort it
day1 = filter(data_clean, num == 1)
day1 = arrange(day1, lon, lat)
day1 = mutate(day1, lon_rank = min_rank(lon), lat_rank = min_rank(lat))
```

Since the temperature points are discrete with respect to the whole space (in the longitude-latitude-axis matrix, only 0.02% of elements are not NA), it is not available to show the United States temperatures by using the image.plot command in the fields package. In this case, we introduce the following method to plot the discrete points' temperature distribution image.

```
library(ggplot2)
library(maps)

statesMap <- map_data("state")

# Extract the states in the main US territory
states = unique(statesMap$region)
# match locations in day1 with regions in statesMap
tempdata=filter(day1, lat < 50 & lat > 25 & lon > (-125) & lon < (-65))

ggplot(data = statesMap) +
  geom_polygon(aes(x = long, y = lat, group = group), color = "black", fill="white") +
  coord_fixed(1.3) +
  guides(fill=FALSE)+  # do this to leave off the color legend
  geom_point(mapping=aes(lon,lat, colour = temp_mean), data = tempdata, size = 0.7) +
  scale_colour_gradient2(low = "blue", high = "red", mid="lightblue", midpoint = 32)
```

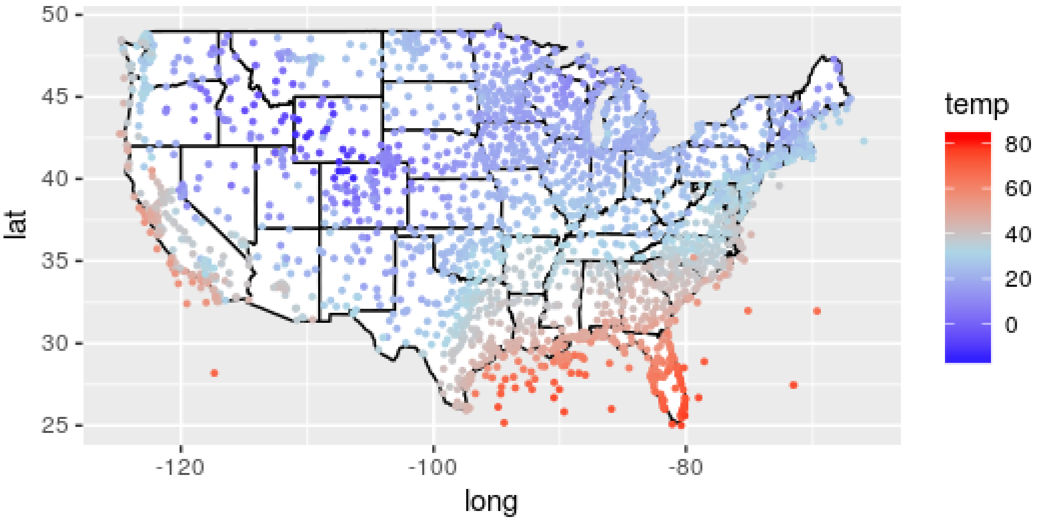

Figure 1: Temperature data of the United States observed on January 1 in 2015

A pronounced temperature gradient is visible from highs of over 85.6 degrees Celsius in the north of the study area which is near to the equator to a low of -14.7 degrees Celsius towards the southern boundary. This is not only indicative of spatial correlation in the dataset, but it also shows that the data are not stationary, as the mean temperature must vary strongly with latitude. A more precise conclusion can be reached by combining Figure 1 with Figure 2.

According to Figure 1 and Figure 2, we can find that on January 1 in 2015, the temperature in most parts of the United States was concentrated at 30 degrees Celsius and extreme temperatures only account for a small proportion.

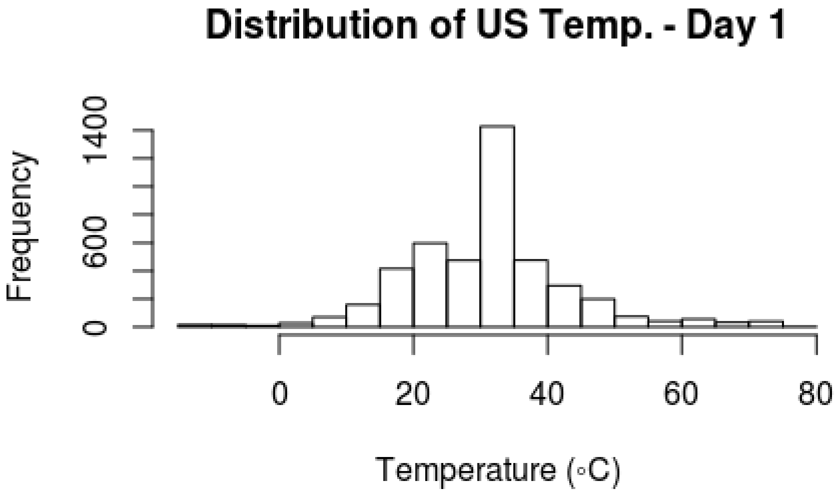

Figure 2: Histogram of US temperature on January 1 in 2015

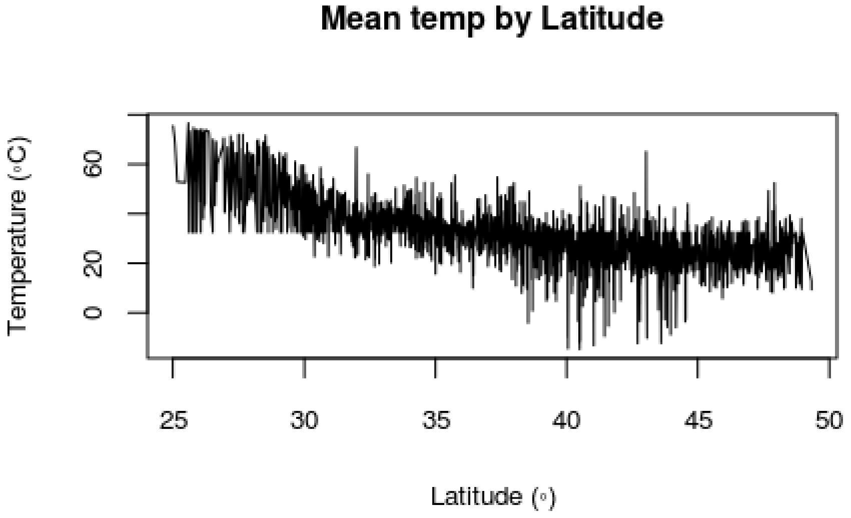

Figure 3: Mean of US temperature in latitude on January 1 in 2015

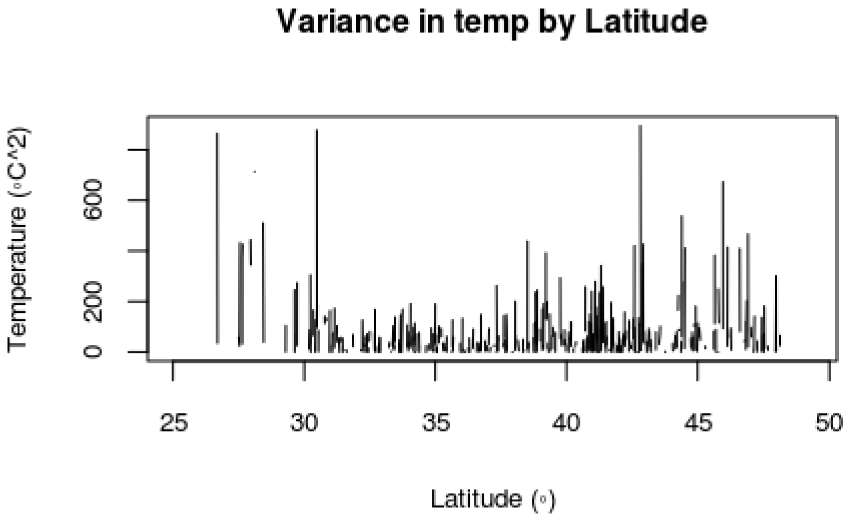

Figure 4: Variance of US temperature in latitude on January 1 in 2015

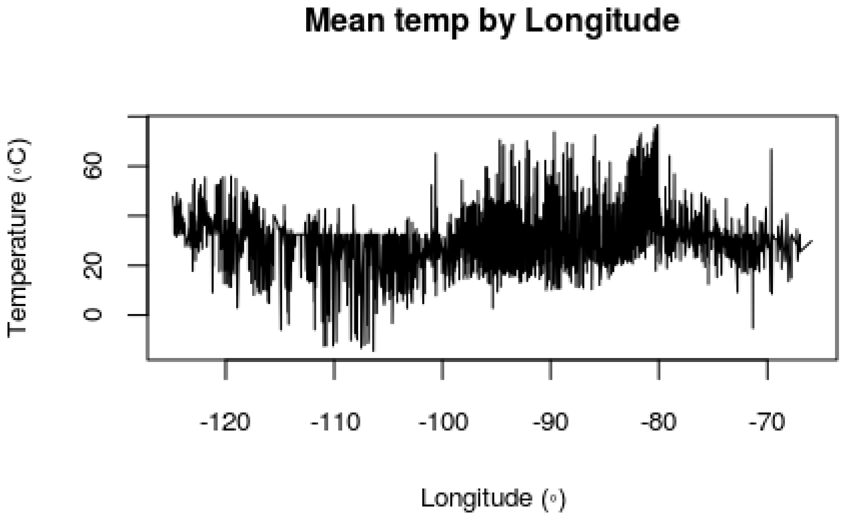

Figure 5: Mean of US temperature in longitude on January 1 in 2015

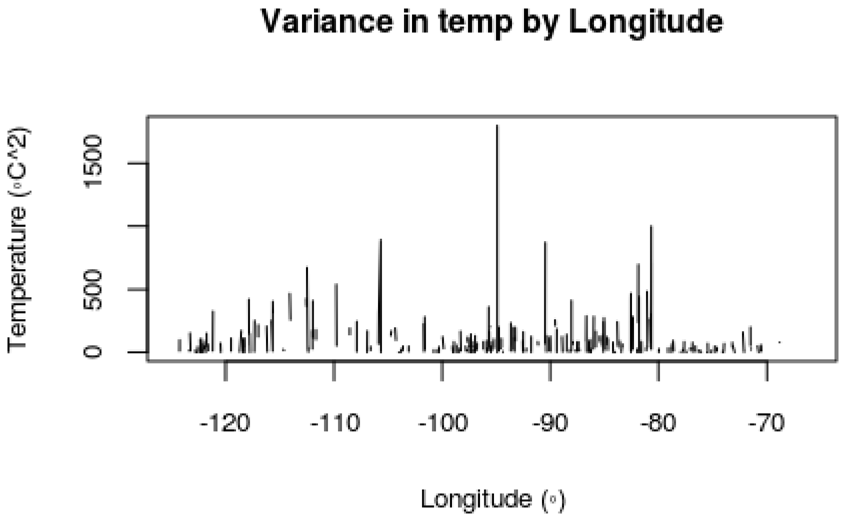

Figure 6: Variance of US temperature in longitude on January 1 in 2015

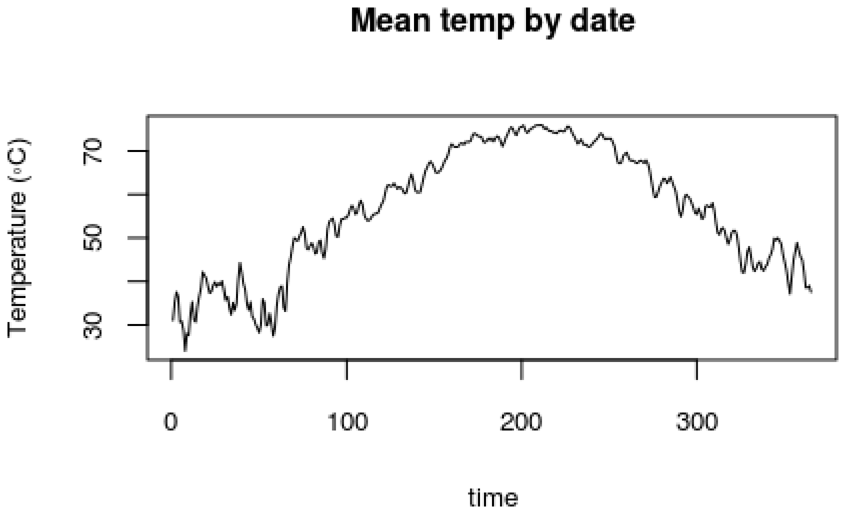

Figure 7: Mean of US temperature by date

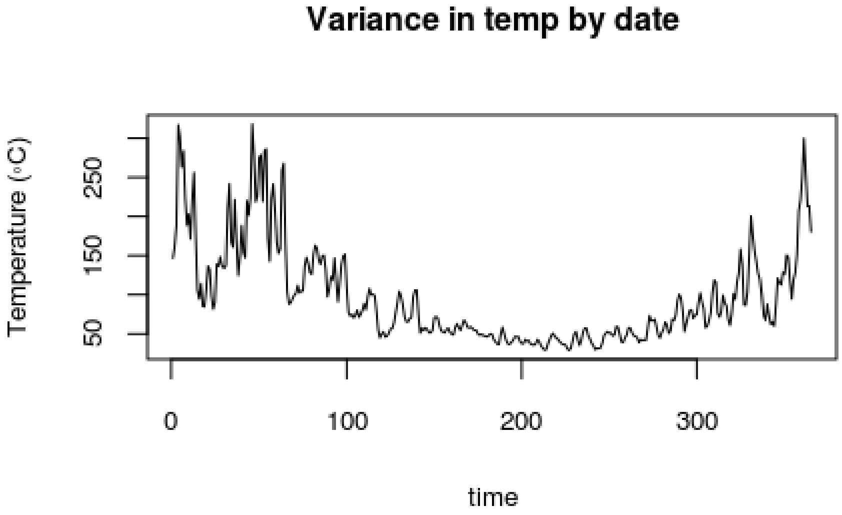

Figure 8: Variance of US temperature by date

Research objectives

In light to Figure 3 - Figure 8, we can confirm that there are some trends in latitude, longitude and date marginally. For simplicity, in variogram modeling and kriging throughout this paper, we will treat the latitude and longitude coordinates as if they are Cartesian. In geostatistics, we generally consider the temperature is normally distributed with mean and covariance structure being functions of latitude, longitude and time. Furthermore, the mean function may be the linear combination or other nonlinear forms of three factors. Apart from this, the covariance may introduce the Matern class, which is popular in spatial statistics. Considering these cases, we will figure out the structure of the normal distribution by the following steps.

1. Construct spatial data analysis regardless of the influence of time and find the most appropriate model.
2. Construct Spatial-Temporal analysis.
3. Various forms of Kriging can be used to attempt to fill gaps caused by orbital clipping and cloud cover so that we can obtain the complete temperature variation diagram across the United States.