

# Exploration and Estimation of North American Climatological Data

James A. Shine  
George Mason University, and  
U.S. Army Topographic Engineering Center  
7701 Telegraph Road  
Alexandria, VA 22315-3864  
jshine1@osf1.gmu.edu

Paul F. Krause  
U.S. Army Topographic Engineering Center  
7701 Telegraph Road  
Alexandria, VA 22315-3864

## ABSTRACT

The availability of spatial and temporal earth data is increasing; for example, NASA's Earth Observing System (EOS) will soon be producing a terabyte of earth data per day. This data will permit more detailed exploration and analysis of earth systems than was possible in the past. One application of particular interest to the authors is the estimation of contour surfaces from point values and the visualization of these surfaces in map form. The authors explored a multivariate data set of approximately 6,000 points in North America and other global locations. Each point contains climatological and other information such as temperature, elevation and precipitation. Spatial correlation was modeled using a semivariogram and several estimation approaches were used to create estimated surfaces and resulting contours. Maps of these results and comparisons between different approaches will be presented.

## INTRODUCTION

### Background:

Weather observations can be considered as data values of various parameters such as temperature, precipitation, evaporation, wind, etc. Climate may be defined as one or more statistics describing the distribution of weather parameters.[1] Climatology can be descriptive such as averages of parameters over a period of time (for example, a year) and can also be predictive (for example, weather models). [2,3] Climate data is frequently collected from weather stations. [4] Both descriptive and predictive applications of climatology require estimates for surface areas, while actual data observations are discrete. Thus estimating values between data observations is essential, as in many other applications of spatial statistics.

Estimation approaches have moved in recent years from classical approaches such as maximum likelihood, discriminant analysis and principal component analysis towards more computer-intensive methods.

### Recent approaches:

Some approaches examined in recent years include:

1. Inverse distance weighting. Points on a surface are calculated by taking a weighted average of nearby data observations. Observations which are closer to

the estimated point are given a larger weight than those that are farther away. Typically the weight is the inverse of the distance or the inverse of the distance squared.

2. Global interpolation. A polynomial surface is fitted by least squares approaches using the data observations; other points on the surface are estimated using the resulting polynomial model.
3. PRISM (Parameter-elevation Regressions on Independent Slopes Model). Parameters such as precipitation and temperature are estimated using elevation values interactively in a knowledge-based approach. PRISM is most effective in mountainous regions. [5]
4. Geostatistical approaches. Spatially autocorrelated data can be modeled by plotting data variation as a function of distance between data observations. The resulting model can then be used to estimate surface points. This method will be described in more detail later.

### DATA SET DESCRIPTION

The authors explored a multivariate data set of approximately 6,000 points in North America and other global locations. The data was collected from World Meteorological Organization (WMO) weather stations during the period 1961 to 1990. Each station provided most or all of 7 different parameters:

1. Mean annual temperature in degrees Fahrenheit.
2. Mean annual precipitation in inches.
3. Continentality index,  

$$K = [1.7A / \sin(L + 10)] - 14,$$
 where A is annual amplitude of temperature in degrees Fahrenheit, and L is the latitude in decimal degrees.
4. Precipitation Effectiveness index,  

$$PE = \sum(1:12) \{ 115 [p / (t - 10)]^{10/9} \}$$
 where p and t are monthly precipitation and average monthly temperature respectively.
5. Potential Evapotranspiration index,  

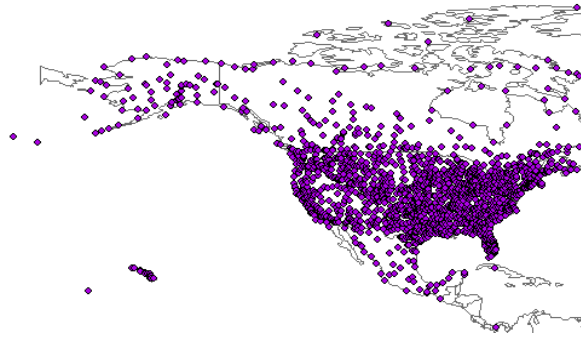
$$PEVAP = \sum(1:12) \{ [10t/I]^{AA} \}$$
 where  

$$AA = 0.4923 + (.00172)I - (.0000771)I^2 + (.000000675)I^3$$

$$I = \sum(1:12) (t / [5^{1.514}])$$
 and t is average monthly temperature.
6. Moisture index,  

$$MI = [(P / PEVAP) - 1] * 100$$
 Where P is annual precipitation in millimeters
7. Elevation, in meters.

For our analysis we used a subset of 1749 data points in North America and surrounding islands. The points in this data set are shown in Figure 1.



**Figure 1:** Map of North American data points

### GEOSTATISTICAL APPROACHES

#### Spatial analysis:

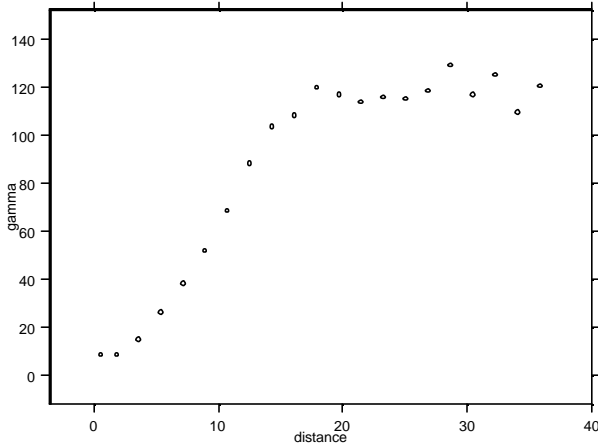
Spatial data (data observations with spatial coordinates, such as latitude and longitude) frequently exhibit spatially dependence; variation between points is not independent of the distance between those points, but instead is a function of this distance. Geostatistical approaches model this dependency for a set of data observations and use the model to predict data values at other locations. This approach dates from the 1950s and 1960s in limited areas such as mining but has recently become applicable to larger data sets with computational advances [6,7,8].

Two main assumptions of geostatistical modeling are stationarity and isotropy. Under stationarity, the variance between two data observations depends on the distance between the observations, as does the correlation, but at a given distance h these two values should be the same regardless of location. The mean is also the same regardless of location. Isotropy means that data statistics are independent of direction; the mean, variance and correlation between observations does not depend on whether the observations are oriented north-south, east-west, or some other direction.

These ideal assumptions are often not met with real data. Some assumption relaxations frequently applied are the Intrinsic Hypothesis (where the variance may be unbounded) and quasi-stationarity (where stationarity applies in a neighborhood of the data, but not in the entire data domain).

The first step in spatial data analysis is the computation and plotting of a variogram. All the observations in a data set are compared to all other observations, and for each distance h between observations, a semivariance function

$\gamma(h)$  is computed which is half of the traditional variance for those observations. A typical variogram is shown in Figure 2.



**Figure 2:** A typical variogram.

Nonspatial, stochastic variation shows up as a baseline or "nugget" variance which is the value of  $\gamma$  at  $h=0$ . Spatial variation shows up in the values of  $\gamma$  as  $h$  increases. A common pattern for variograms is for  $\gamma$  to increase up to certain values of  $h$ , and then remain nearly constant or only increase slightly for higher values of  $h$ . The value of  $h$  where  $\gamma$  no longer increases is called the range, and the value of  $\gamma$  at this  $h$  is called the sill.

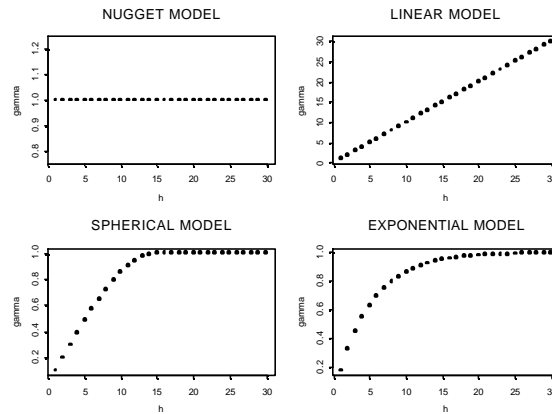
In nonspatial analysis a set of data observations is often compared to various theoretical distributions such as normal, exponential, gamma, etc. The distribution that most closely matches the data is chosen as a model for inference and estimation. A similar approach is followed in spatial analysis. The variogram, which is a statistic of the data observations, is compared with theoretical variogram models and the one with the best fit is chosen. This model can then be used for estimation.

Some common models are listed below. The left side of the equation in all cases is assumed to be  $\gamma(h)$  for simplicity.

1. nugget effect model:  
 $1$  at all  $h$  except  $0$   
 (this model assumes no spatial correlation)
2. power model:  
 $h^w$ ,  $0 < w < 2$   
 (if  $w=1$ , this is a linear model)
3. spherical model:  
 $(1.5h/a) - 0.5(h/a)^3$  if  $h \leq a$ ,  
 $1$  otherwise

4. exponential model:  
 $1 - \exp(-3h/a)$

Some plots of the shape of these functions are shown in Figure 3.



**Figure 3:** Plots of four common variogram models: nugget effect, linear, spherical ( $a=15$ ), and exponential ( $a=15$ ).

If we compare the empirical variogram in Figure 1 with the model variograms in Figure 2, it is evident that either a spherical model or an exponential model will provide the best fit.

**Kriging approaches:**

Kriging is a weighted average interpolation approach which uses a variogram model to determine the weights. A small amount of neighboring points are selected and the model is then applied. A matrix inversion is involved.[9]

Cokriging is a multivariate form of kriging. [10] Disjunctive kriging involves the use of indicator functions. [11]

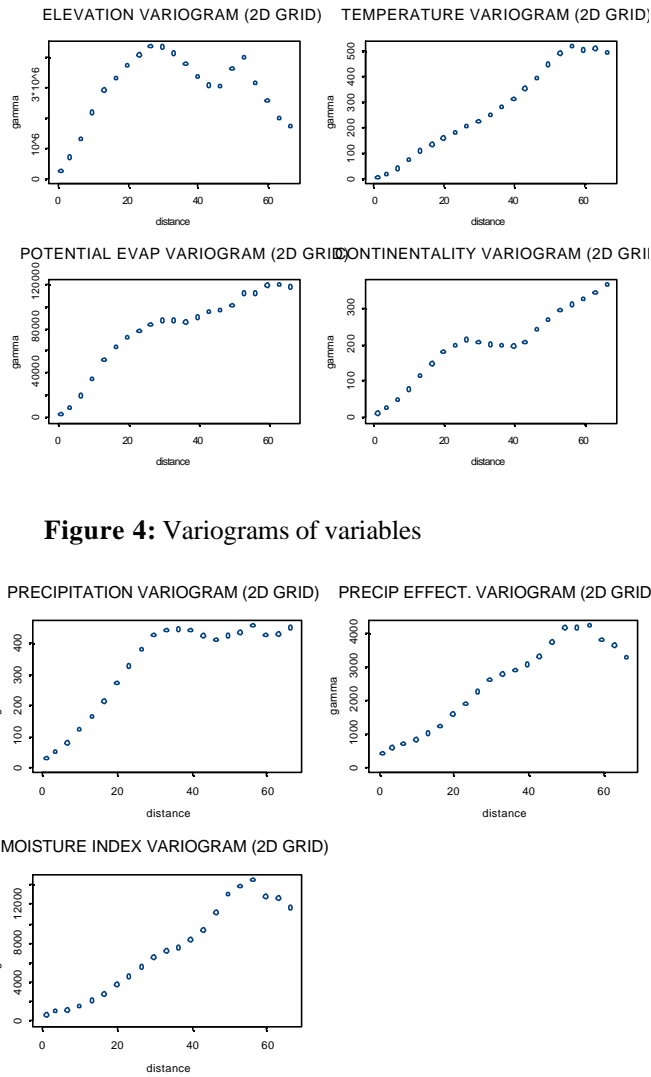
Other kriging approaches are available but are not discussed here.

**ANALYSIS**

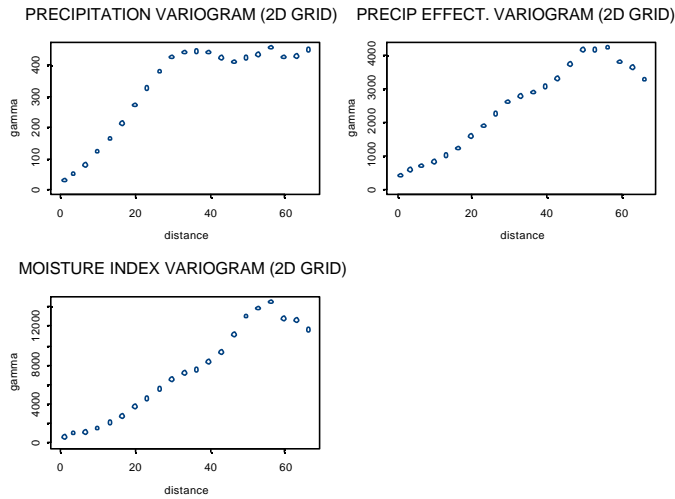
Much of the analysis presented here was performed using the Geostatistical Analyst module for ArcInfo 8 from ESRI [12]. The principal author was involved in beta testing for this product, which should be available in the fall of 2000.

ESRI agreed to permit disclosure of the results shown here provided that no evaluation of the software was presented.

Exploratory analysis of the seven data sets revealed a strong north-south trend and non-normal distributions. Variograms of the seven variables definitely show spatial dependence and seem to support exponential or spherical models (Figures 4 and 5); anisotropy does not appear to be a problem.



**Figure 4:** Variograms of variables



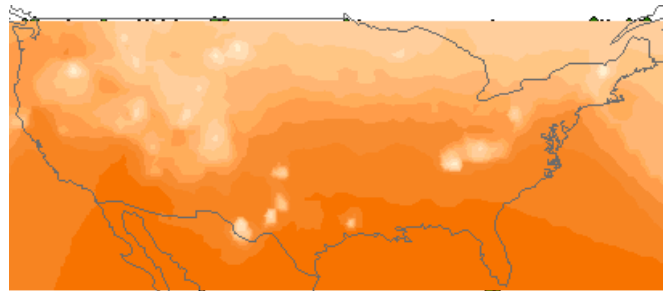
**Figure 5:** More variograms of variables

These variograms are inherently distorted since they are computed on latitude and longitude as a 2-dimensional grid and the observations are actually on a roughly spherical surface (the Earth), so the distances between observations

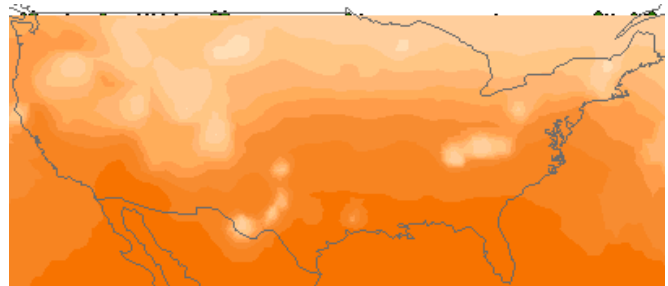
which are not near each other will be inaccurate. [13] Code has been written to compute accurate variograms using projections from the latitude and longitude.

Surface maps for three of the variables (elevation, precipitation, and temperature) were created using several of the techniques discussed earlier: inverse distance weighting, ordinary kriging, and cokriging. Surfaces created from disjunctive kriging were looked at briefly but were computationally very slow and were left out of the extensive analysis. Accuracy comparisons were performed against surface maps developed by the U.S. Geological Survey (USGS) and the National Climate Data Center (NCDC). [14,15] Both continental maps and maps restricted to the continental United States were compared; the latter were created by deleting other North American data points to facilitate accuracy assessment with the USGS and NCDC data.

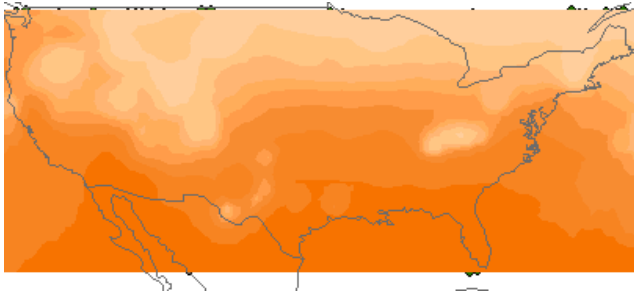
Three results for the temperature maps are shown in Figures 6, 7, 8 and 9.



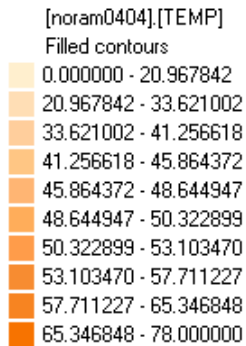
**Figure 6:** Temperature map of continental United States created using Inverse Distance Weighting



**Figure 7:** Temperature map of continental United States created using Ordinary Kriging (Exponential Model)



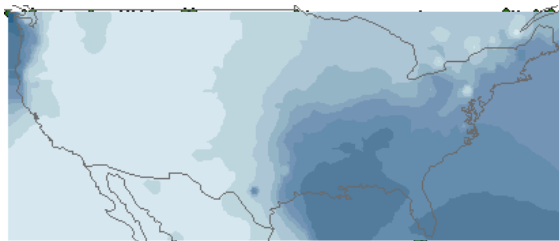
**Figure 8:** Temperature map of continental United States created using Cokriging (Exponential Model, other variables were precipitation, elevation and continentality)



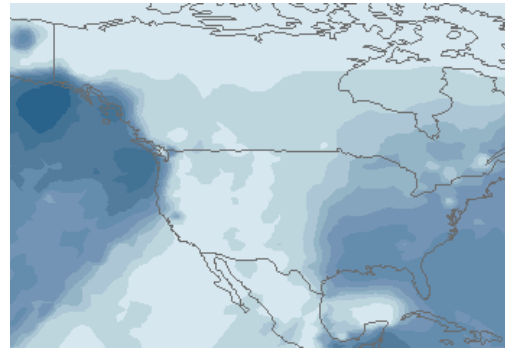
**Figure 9:** Scale of color values for Figures 6,7 and 8.

At this scale there was no statistically significant difference between the three techniques. However, a visual comparison shows a number of artifacts: small areas of low temperature surrounded by areas of higher temperature. Ordinary kriging eliminates some of these artifacts and cokriging eliminates almost all of them.

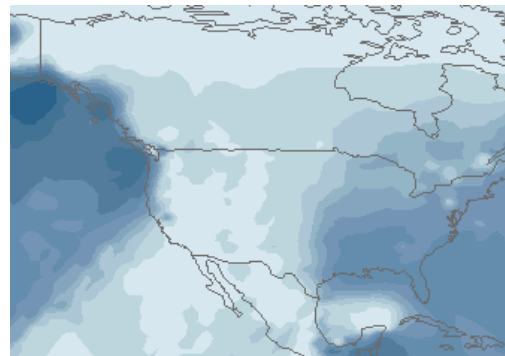
Precipitation results were much smoother; it was hard to tell any difference between the approaches (Figures 10,11, 12 and 13). Again, there was no statistically significant difference.



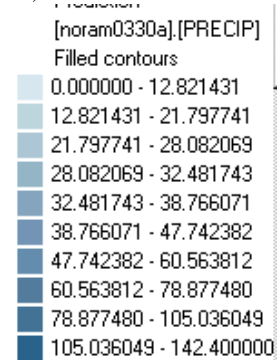
**Figure 10:** Precipitation map of continental United States created using Inverse Distance Weighting



**Figure 11:** Precipitation map of continental United States created using Ordinary Kriging (Exponential Model)

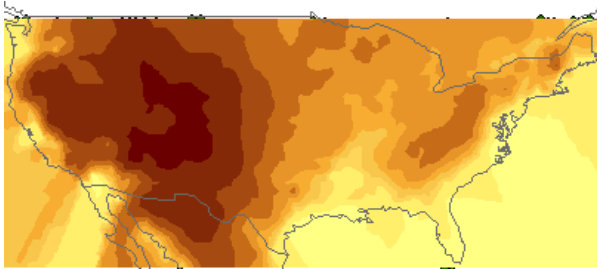


**Figure 12:** Precipitation map of continental United States created using Cokriging (Exponential Model, other variables were precipitation, elevation and continentality)

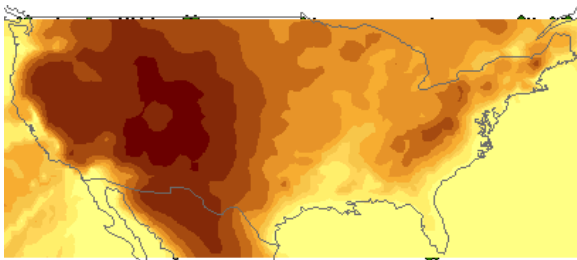


**Figure 13:** Scale of color values for Figures 10,11 and 12

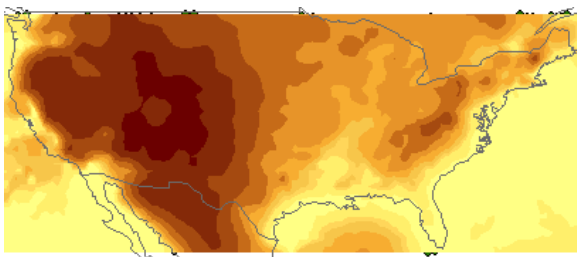
Elevation results were also not statistically significant, and artifacts were present in all 3 approaches (Figures 14,15,16 and 17)



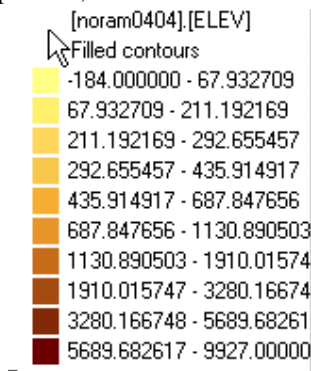
**Figure 14:** Elevation map of continental United States created using Inverse Distance Weighting



**Figure 15:** Elevation map of continental United States created using Ordinary Kriging (Exponential Model)



**Figure 16:** Elevation map of continental United States created using Cokriging (Exponential Model, other variables were precipitation, elevation and continentality)



**Figure 17:** Scale of color values for Figures 14,15 and 16

One problem in all the elevation maps is the existence of positive elevations in the ocean. This is inevitable in a weighted average approach and outside processing is needed to clean up this problem.

## CONCLUSIONS AND FUTURE WORK

The results shown indicate several things. On a continental scale, kriging and cokriging do not perform significantly better than inverse distance weighting. However, both seem to be less vulnerable to created artifacts, with cokriging slightly better than kriging at this. For a rough scale, any of the three approaches is roughly equivalent; for finer scales, cokriging is probably the preferred approach when the data is multivariate. However, cokriging is more computationally intensive than kriging, and both kriging and cokriging are more computationally intensive than inverse distance weighting. This may be a factor when processing large amounts of data and/or large data sets.

Given more time, there are a number of tasks which could be done to expand on the results reported here. More quantitative results for comparing different map surfaces against known maps and against each other would probably be the most useful future effort. Recomputing the variograms, variogram models, and kriged surfaces using 3-dimensional projections to compute distances would also be desirable. Obtaining data for Alaska and western Canada would permit a larger area of land to be compared. Looking at detrended and transformed kriging approaches might also yield interesting results.

## REFERENCES

- [1] Ikeda, S., et.al, "Statistical Climatology", Elsevier, 1980.
- [2] Griffiths, J.F., and Driscoll, D.M., "Survey of Climatology", Merrill Publishing, 1982.
- [3] Stringer, E.T., "Techniques of Climatology", W.H. Freeman, 1972.
- [4] Lydolph, P.E., "The Climate of the Earth", Rowman and Allanheld, 1985.
- [5] Daly, C., Neilson, R.P., and Phillips, D.L., "A statistical-topographic model for mapping climatological precipitation over mountainous terrain", Journal of Applied Meteorology, Volume 33, pp.140-158, 1994.
- [6] Matheron, G., "The Theory of Regionalized Variables and Its Applications", Fontainebleau, 1971.
- [7] Oliver, M.A., and Webster, R., "Spatial Analysis of Part of Fort Benning Using Remote Imagery", US Army Environmental Research Office, London, 1997.

- [8] Oliver, M.A. and Webster, R., "Geostatistical Analysis of High Resolution Multispectral Imagery", US Army Environmental Research Office, London, 1998.
- [9] Goovaerts, P., "Geostatistics for Natural Resources Evaluation", Oxford University Press, 1997.
- [10] Wackernagel, H., "Multivariate Geostatistics", Springer, 1998.
- [11] Rivoirard, J., "Introduction to Disjunctive Kriging and Non-Linear Geostatistics", Clarendon Press, 1994.
- [12] Environmental Systems Research Institute Inc., "Documentation for Geostatistical Analyst", 2000.
- [13] MacEachren, A.M., "Some Truth with Maps: A Primer on Symbolization and Design", Association of American Geographers, 1994.
- [14] Visher, S.S., "Climatic Atlas of the United States", Harvard University Press, 1954.
- [15] U.S. Department of the Interior, "The National Atlas of the United States of America", U.S. Geological Survey, 1970.