MODELING MULTIVARIATE SPATIAL DATA

I. Research Aim

This aim of research is to build a model to predict the atmospheric pressure based on special covariates. The main constrains in this case are the dependencies between time series data as the pressure are measured daily and spatial data since the measurements are made on different weather stations locations.

II. Datasets

There are two data sets used in this research. The first one is called Weather data. It consists of daily average pressure readings (in inches) from May 2, 2017 to September 30, 2017 (152 days) measured by 17 stations in Oklahoma State, USA.

	YEAR	MONTH	DAY	ACME	CHAN	CHIC	ELRE	FTCB	GUTH	HINT	KETC	MARE	NINN	OKCE	PAUL	PERK	SHAW	SPEN	WASH	WATO
1	2017	5	2	28.50	28.90	28.74	28.45	28.42	28.76	28.19	28.69	28.78	28.64	28.67	28.88	28.90	28.77	28.61	28.69	28.13
2	2017	5	3	28.54	28.90	28.78	28.50	28.48	28.79	28.26	28.71	28.80	28.68	28.69	28.87	28.91	28.76	28.63	28.71	28.20
3	2017	5	4	28.70	29.03	28.93	28.63	28.63	28.91	28.39	28.88	28.92	28.84	28.83	29.04	29.03	28.91	28.76	28.87	28.31
4	2017	5	5	28.66	29.00	28.89	28.58	28.58	28.87	28.34	28.84	28.88	28.79	28.79	29.01	28.99	28.88	28.73	28.83	28.26
5	2017	5	6	28.57	28.92	28.80	28.49	28.48	28.78	28.24	28.76	28.79	28.70	28.71	28.93	28.91	28.80	28.64	28.75	28.16

The second one is the corresponding information of each stations. In terms of its Station Number (STNM), Station Identifier (STID), Station Name (NAME), Nearest Incorporated Town (CITY), Range From Town To Station (RANG), Compass Direction From Town To Station (CDIR), Country (CNTY), North Latitude (NLAT), East Longitude (ELON), and Elevation In Meters (ELEV).

	stnm	stid	name	city	rang	cdir	cnty	nlat	elon	elev
1	110	ACME	Acme	Rush Springs	4	WNW	Grady	34.80833	-98.02325	397
2	24	CHAN	Chandler	Sparks	3	NNE	Lincoln	35.65282	-96.80407	291
3	27	CHIC	Chickasha	Chickasha	2	SSE	Grady	35.03236	-97.91446	328
4	34	ELRE	El Reno	El Reno	5	WNW	Canadian	35.54848	-98.03654	419
5	40	FTCB	Fort Cobb	Fort Cobb	4	NNW	Caddo	35.14887	-98.46607	422
6	43	GUTH	Guthrie	Guthrie	4	WSW	Logan	35.84891	-97.47978	330
7	45	HINT	Hinton	Hinton	7	W	Caddo	35.48439	-98.48151	493
8	53	KETC	Ketchum Ranch	Velma	7	NW	Stephens	34.52887	-97.76484	341
9	59	MARE	Marena	Coyle	7	N	Payne	36.06434	-97.21271	327
10	109	NINN	Ninnekah	Ninnekah	2	NNW	Grady	34.96774	-97.95202	356
11	130	OKCE	Oklahoma City East	Oklahoma City	3	E	Oklahoma	35.47226	-97.46414	355
12	74	PAUL	Pauls Valley	Pauls Valley	1	SSW	Garvin	34.71550	-97.22924	291
13	76	PERK	Perkins	Perkins	2	NNW	Payne	35.99865	-97.04831	292
14	84	SHAW	Shawnee	Shawnee	3	NNW	Pottawatomie	35.36492	-96.94822	328
15	87	SPEN	Spencer	Spencer	2	ENE	Oklahoma	35.54208	-97.34146	373
16	99	WASH	Washington	Washington	6	SSW	McClain	34.98224	-97.52109	345
17	100	WATO	Watonga	Watonga	7	W	Blaine	35.84185	-98.52615	517

III. Methodology

1. Predictive Model Building
   There are two approaches to build the model that can be done to handle the dependencies between serial and spatial effects.

   First Approach

   1. Build a single model for each station using the lag variable(s) as covariate(s)

$$\begin{pmatrix} y_{i,2} \\ \vdots \\ y_{i,151} \end{pmatrix} = \alpha_i + \beta \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,150} \end{pmatrix} i = 1, \dots, 17$$

2. Build a model using of the intercepts of each model as resulted from the previous step with the spatial covariates

$$\begin{pmatrix} \alpha_1 \\ \vdots \\ \alpha_{17} \end{pmatrix} = \gamma_0 + \gamma_1 \begin{pmatrix} LAT_1 \\ \vdots \\ LAT_{17} \end{pmatrix} + \gamma_2 \begin{pmatrix} LON_1 \\ \vdots \\ LON_{17} \end{pmatrix} + \gamma_3 \begin{pmatrix} ELEV_1 \\ \vdots \\ ELEV_{17} \end{pmatrix}$$

3. Compute the predicted value for each station using the model as resulted from step (2) as the new intercepts combined with the lag variable(s) used in step (1)

$$\begin{pmatrix} \hat{y}_{i,2} \\ \vdots \\ \hat{y}_{i,151} \end{pmatrix} = \hat{\alpha}_i + \beta \begin{pmatrix} y_{i,1} \\ \vdots \\ y_{i,150} \end{pmatrix} i = 1, \dots, 17$$

Second Approach

1. Build a single model for all stations at once using the lag variable(s) and spatial variables as covariates.

$$\begin{pmatrix} y_{1,2} \\ \vdots \\ y_{1,151} \\ \vdots \\ y_{17,2} \\ \\ y_{17,151} \end{pmatrix} = \alpha + \beta_1 \begin{pmatrix} y_{1,1} \\ \vdots \\ y_{1,150} \\ \vdots \\ y_{17,1} \\ \\ y_{17,150} \end{pmatrix} + \beta_2 \begin{pmatrix} LAT_1 \\ \vdots \\ LAT_1 \\ \vdots \\ LAT_{17} \\ \\ LAT_{17} \end{pmatrix} + \beta_3 \begin{pmatrix} LON_1 \\ \vdots \\ LON_1 \\ \vdots \\ LON_{17} \\ \\ LON_{17} \end{pmatrix} + \beta_4 \begin{pmatrix} ELEV_1 \\ \vdots \\ ELEV_1 \\ \vdots \\ ELEV_{17} \\ \\ ELEV_{17} \end{pmatrix}$$

2. Compute the predicted value for each station directly from the model as resulted from previous step.

2. Residuals Exploration
The standardized residuals of each station can be computed after the best fitted model is achieved thus they will have zero mean and unit variance. The standardized residuals for all stations are assumed have joint normal distribution. Exploratory data analysis can be done with standardized residuals through histograms and scatter plots.

3. Parameters Estimate for Residuals Covariance Matrix
The covariance matrix between standardized residuals of the 17 stations is assumed follow multivariate normal distribution with vector mean 0 and covariance matrix $\sum$ which defined based on the spatial distance between stations formulated as follow:

$$\sum_{i,j} = e^{-r\{dist_{i,j}\}^{\alpha}} \text{ where } r > 0 \text{ and } 0 < \alpha \leq 2$$

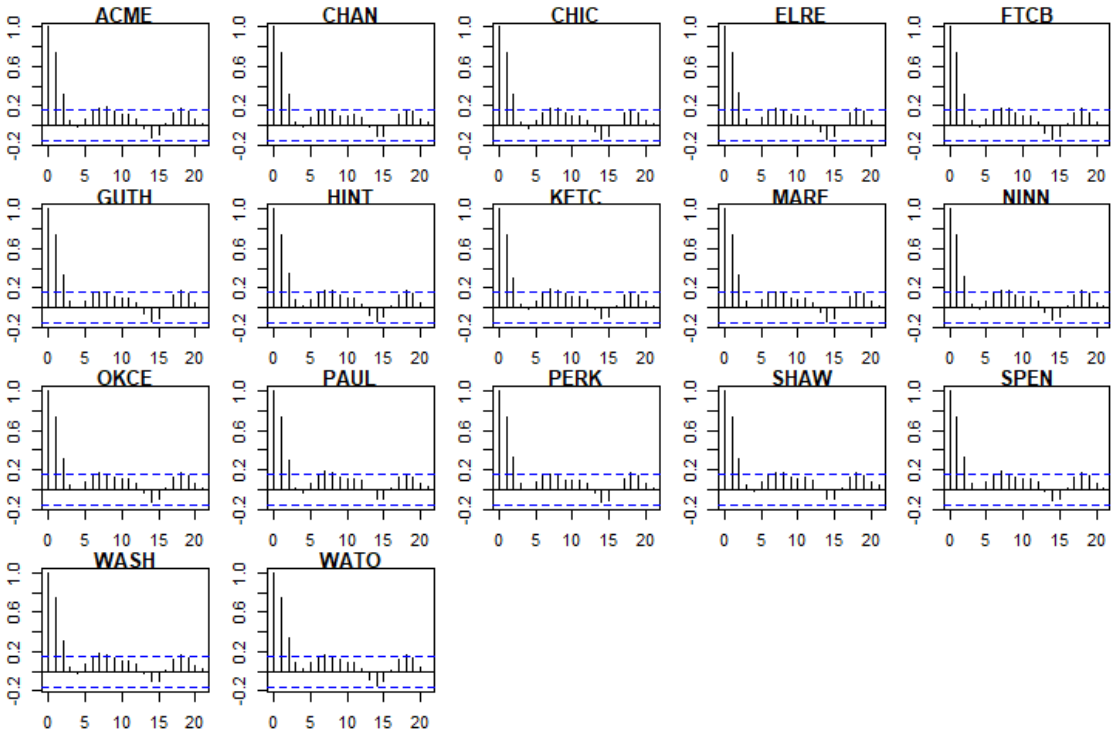
The estimation of parameter r and α is computed through Maximum Likelihood Estimation

4. Confidence Intervals for Parameters Estimate
   The confidence interval for parameter r and α is achieved through bootstrap method.

IV. RESULTS

1. Predictive Model Building

Since there are two approaches to build the model in this case. The first step in predictive model building is compare the result of both approaches. Before evaluating both approaches, below is the ACF plot of the pressure data for each station. According to this plot, it is obviously seen that the pressure data of each station has serial correlation. This evaluation will be useful to define the appropriate model for this case, which has independent residuals or no autocorrelation.



1.1. First Approach

The first step of this approach is building a single linear model using 1 lagged variable as the covariates. This step resulted in 17 linear model with each has a pair of intercepts and covariates. Since it is assumed that the intercepts depend on spatial covariates, the next step is to make a new model using the 17 intercepts as a response variable and latitude, longitude, and elevation as the spatial covariates. Below is the R summary of this model.

```
Call:
lm(formula = alpha ~ lat + lon + elev)

Residuals:
      Min        1Q    Median        3Q       Max
-0.043278 -0.021525  0.001473  0.020049  0.053117

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -1.1753069  3.6508994  -0.322   0.7526
lat          0.0507501  0.0201596   2.517   0.0257 *
lon         -0.0712844  0.0335144  -2.127   0.0531 .
elev        -0.0015058  0.0002703  -5.570 9.07e-05 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02986 on 13 degrees of freedom
Multiple R-squared:   0.86,     Adjusted R-squared:  0.8277
F-statistic: 26.62 on 3 and 13 DF,  p-value: 8.024e-06
```

Through this model then the predicted intercept can be computed and then plugged in to the 17 models to count the predicted response. The following table shows the parameter estimates of models for each station, the predicted intercept as resulted from the second step model, the standardized residuals mean and standard deviation for each station model.

	intercept	beta1	pred_int	mean_stdres	sd_stdres
1	6.944985	0.7571122	6.980965	-1.876334e-17	0.9712779
2	7.127218	0.7538288	7.096525	-2.379931e-17	0.9801075
3	7.069860	0.7546694	7.088477	5.661725e-18	0.9759426
4	6.975191	0.7554134	6.986349	2.071562e-18	0.9784684
5	7.022358	0.7536556	6.992170	3.946343e-17	0.9763493
6	7.052642	0.7551712	7.095920	-4.127292e-17	0.9806513
7	6.881537	0.7566151	6.903390	-4.889303e-17	0.9780233
8	7.085802	0.7537481	7.032685	4.039910e-17	0.9751813
9	7.100717	0.7536283	7.092332	-1.570208e-18	0.9856573
10	7.066345	0.7539606	7.045714	1.639730e-17	0.9752295
11	7.055029	0.7544611	7.038046	-2.364333e-17	0.9794221
12	7.057740	0.7561996	7.079264	-1.531184e-17	0.9717217
13	7.139440	0.7533151	7.129981	2.029953e-17	0.9846202
14	7.029770	0.7561611	7.036477	2.375929e-17	0.9768738
15	7.007213	0.7555988	7.005741	1.688340e-17	0.9773601
16	7.000435	0.7566786	7.032295	-3.673561e-17	0.9737428
17	6.908624	0.7549776	6.888575	2.569243e-17	0.9810207

1.2. Second Approach

Compared to the first approach, the second approach is less complex since there is only one model needed. The following result shows all the coefficient estimate using one lagged variable and three spatial covariates. The overall R-Squared of this model is approximately 89.8%

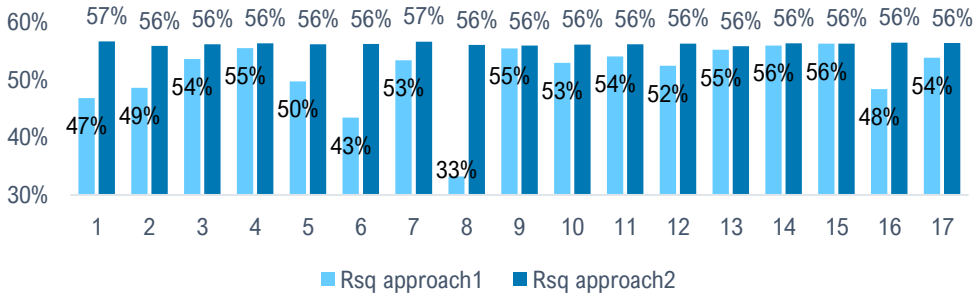
```
Call:
lm(formula = yy ~ yylag1 + LAT + LON + ELEV)

Residuals:
      Min        1Q    Median        3Q       Max
-0.255034 -0.042000 -0.002603  0.049037  0.275794

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  7.755e+00  8.769e-01   8.844   <2e-16 ***
yylag1       7.552e-01  1.318e-02  57.314   <2e-16 ***
LAT         -1.439e-03  4.259e-03  -0.338    0.736
LON          4.006e-03  7.083e-03   0.566    0.572
ELEV        -7.853e-04  7.104e-05 -11.054   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.07752 on 2562 degrees of freedom
Multiple R-squared:  0.8984,    Adjusted R-squared:  0.8983
F-statistic:  5666 on 4 and 2562 DF,  p-value: < 2.2e-16
```

1.3. First Approach and Second Approach Comparison

The R-squared of model from the first approach can be computed through the same way as the second approach which combine all the actual and predictive responses from all stations. According to the overall R-squares, the second approach is slightly higher than the first approach by around 1.2%.

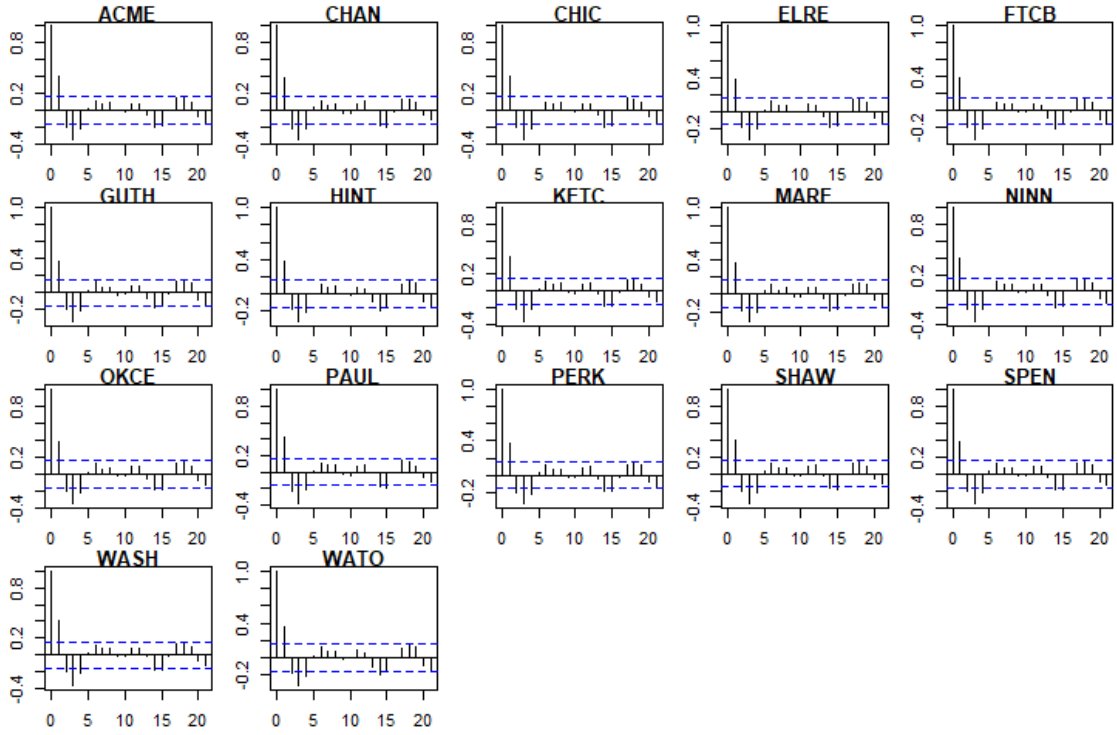
Approach	R-Squared
1	0.8868937
2	0.8984327

After re-compute R-squared based on the actual and predictive responses of each stations, it is clearly seen that the model as resulted from the second approach returns higher R-squared for all stations. Therefore, the predictive model for this case is decided to be built using the second approach.



1.4. Model Improvements

There are several constrains to get the best model. The first one is having the best Goodness of Fit. It can be evaluated from R-Squared and AIC. The second one is having independent Residuals. As shown in the ACF of actual response variable, there is autocorrelation within the data. Thus, the residuals of the appropriate model should be independent or not have autocorrelation. The following graphs are ACF plots of the residuals from the previous model using the second approach. A model with one lagged variable with three spatial covariates.



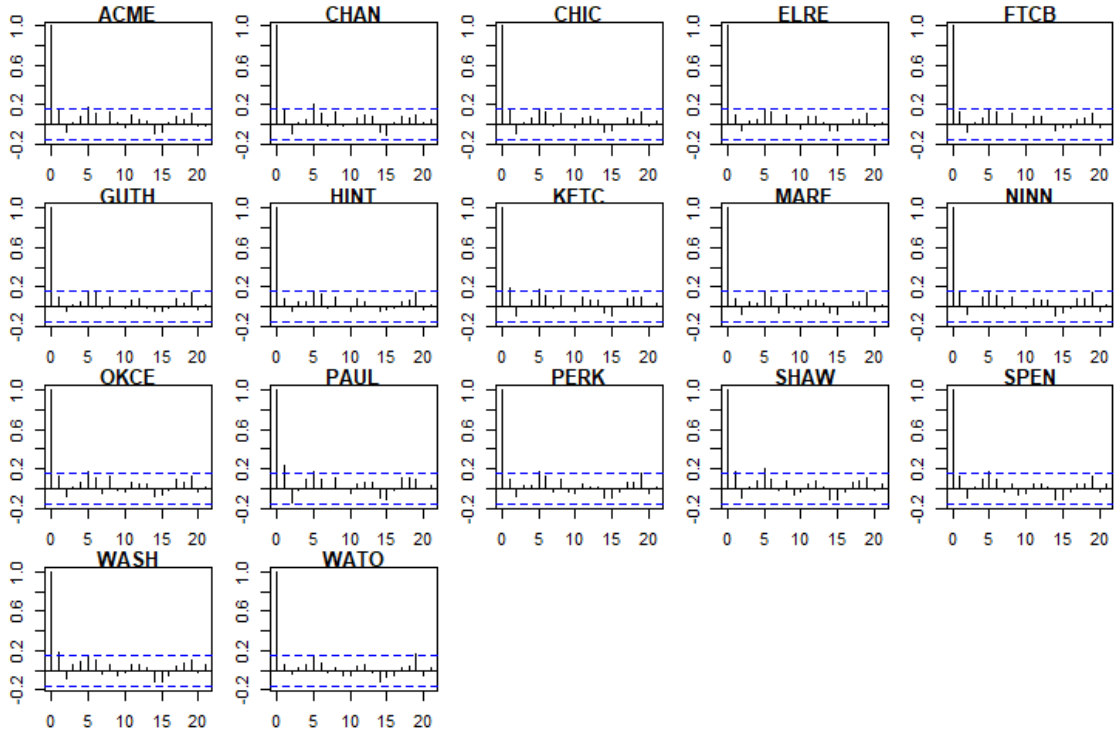
Based on the above ACF plots, the autocorrelation is reduced compared to the original data, but it still exists. Therefore, more lag variable should be included. Below is the result of model with two lagged variables and three spatial covariates.

```
Call:
lm(formula = yy ~ yylag1 + yylag2 + LAT + LON + ELEV)

Residuals:
      Min        1Q    Median        3Q       Max
-0.202275 -0.038045 -0.001281  0.043111  0.233579

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.219e+01  7.659e-01  15.920   <2e-16 ***
yylag1       1.144e+00  1.691e-02  67.673   <2e-16 ***
yylag2      -5.279e-01  1.710e-02 -30.867   <2e-16 ***
LAT         -2.337e-03  3.657e-03  -0.639    0.523
LON          6.701e-03  6.081e-03   1.102    0.271
ELEV        -1.230e-03  6.263e-05 -19.640   <2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06633 on 2544 degrees of freedom
Multiple R-squared:  0.9258,    Adjusted R-squared:  0.9257
F-statistic:  6350 on 5 and 2544 DF,  p-value: < 2.2e-16
```

There is an increase of R-squared by 2.7% compared to the previous model. According to the p-values for each covariate, both latitude and longitude are not significant to the model. The same condition applies in the previous model. Hence, to reduce the complexity of the model, the next step is do forward stepwise selection to the model. As the final model from stepwise selection, both latitude and longitude are excluded from the model. It can be seen in the following result that the stepwise model has the same R-squared with the original model, which means both spatial covariates do not contribute to improve the better fit of the model.

```
Call:
lm(formula = yy ~ yylag1 + yylag2 + ELEV)

Residuals:
      Min        1Q    Median        3Q       Max
-0.202569 -0.038137 -0.001095  0.043156  0.234051

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  1.146e+01  3.633e-01   31.55   <2e-16 ***
yylag1       1.144e+00  1.690e-02   67.70   <2e-16 ***
yylag2      -5.276e-01  1.710e-02  -30.86   <2e-16 ***
ELEV        -1.278e-03  4.534e-05  -28.18   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06632 on 2546 degrees of freedom
Multiple R-squared:  0.9258,    Adjusted R-squared:  0.9257
F-statistic: 1.059e+04 on 3 and 2546 DF,  p-value: < 2.2e-16
```

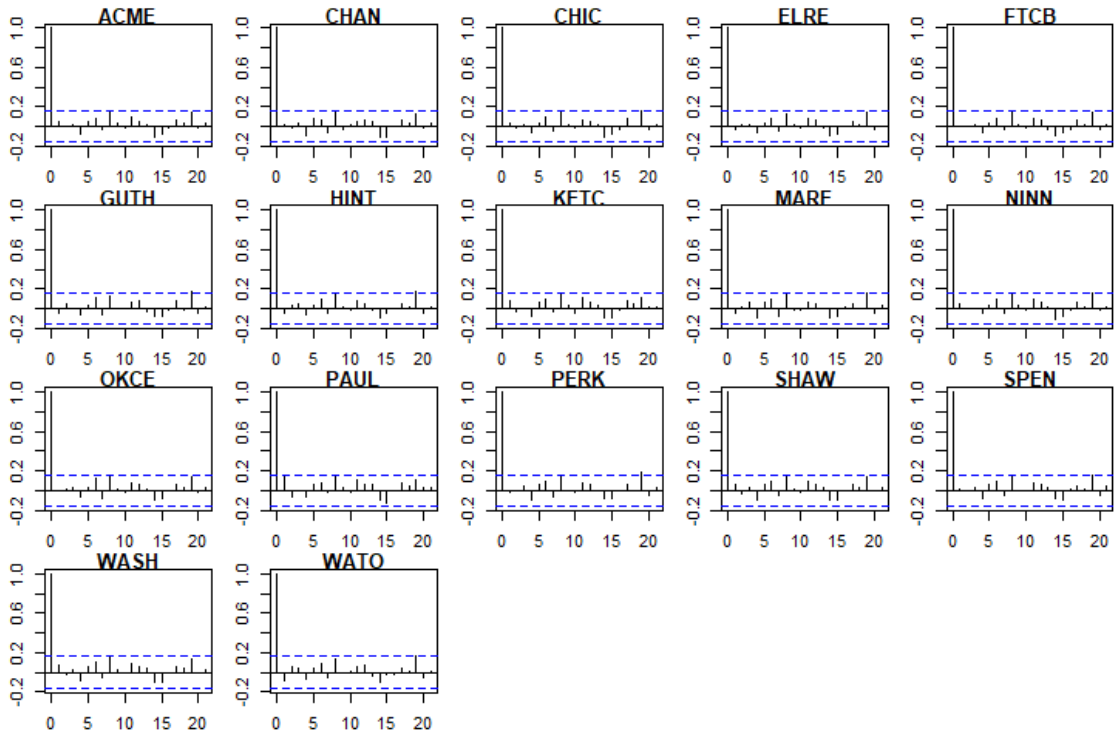
The following graphs are ACF plots of model with two lagged variables and elevation as the covariates.



Based on the above ACF plots, overall the residuals seem nearly independent. However, the autocorrelation still exists for some stations, such as ACME, CHAN, KFTC, OKCE, PAUL, PERK, SHAW, and WASH. Thus, more lag variable is worth to be included. Below is the result of model with three lagged variables and elevation as the covariates.

```
Call:
lm(formula = yy ~ yylag1 + yylag2 + yylag3 + ELEV)

Residuals:
      Min        1Q    Median        3Q       Max
-0.207844 -0.036877  0.001259  0.039616  0.238646

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.594e+00  4.192e-01   20.50   <2e-16 ***
yylag1       1.265e+00  1.916e-02   65.99   <2e-16 ***
yylag2      -7.866e-01  2.732e-02  -28.79   <2e-16 ***
yylag3       2.347e-01  1.939e-02   12.10   <2e-16 ***
ELEV        -9.571e-04  5.073e-05  -18.86   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.06405 on 2528 degrees of freedom
Multiple R-squared:  0.9308,    Adjusted R-squared:  0.9307
F-statistic:  8506 on 4 and 2528 DF,  p-value: < 2.2e-16
```

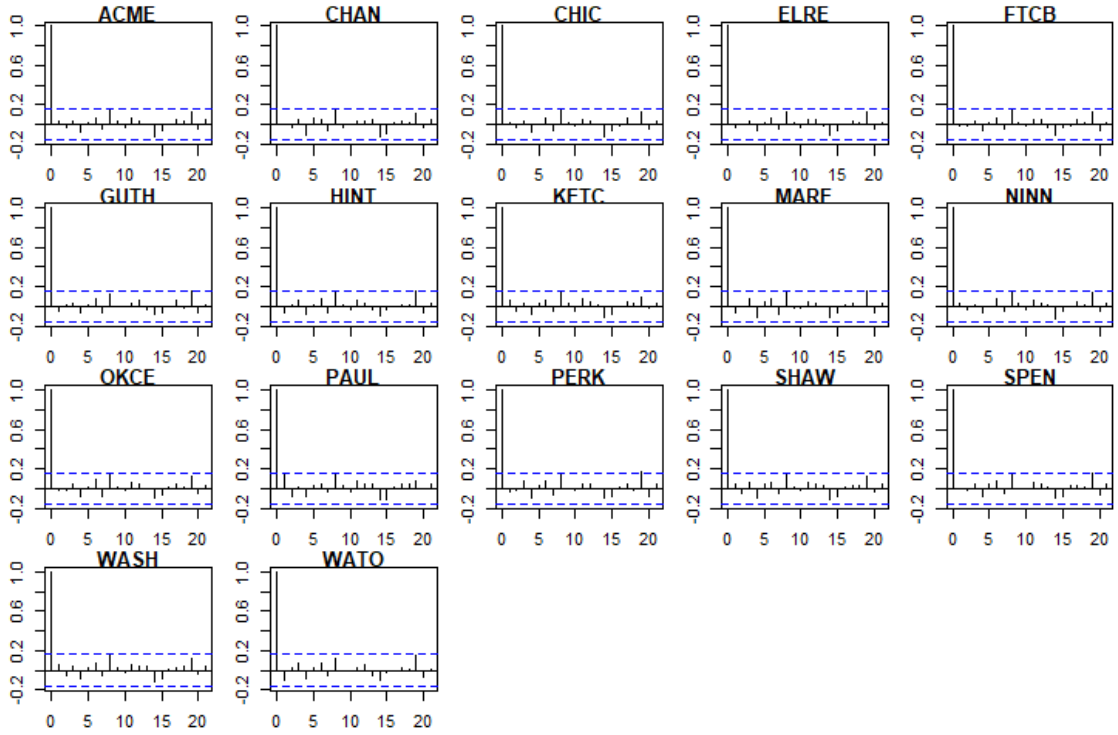
There is a slight increase of R-squared by 0.5% compared to the previous model. The 3 lagged variables and elevation have significant contribution to the model. Furthermore, the following ACF plots also show that the residuals are independent in all stations as most correlations close to zero and not exceed the upper bound for all lag 1 to 20.



This model seems to be the best from all the previous model that have been fitted to the data. However, to make sure that it is the best fit, one more lagged variable can be included to the model. The following result is model with four lagged variables and elevation as covariates.

```
Call:
lm(formula = yy ~ yylag1 + yylag2 + yylag3 + yylag4 + ELEV)

Residuals:
      Min        1Q    Median        3Q       Max
-0.210426 -0.037761  0.000695  0.039059  0.236917

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  8.1182305  0.4575444  17.743  < 2e-16 ***
yylag1       1.2697315  0.0198457  63.980  < 2e-16 ***
yylag2      -0.7704961  0.0315428 -24.427  < 2e-16 ***
yylag3       0.1888967  0.0313978   6.016 2.05e-09 ***
yylag4       0.0404552  0.0199077   2.032   0.0422 *
ELEV        -0.0009045  0.0000547 -16.536  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06387 on 2510 degrees of freedom
Multiple R-squared:  0.9313,    Adjusted R-squared:  0.9312
F-statistic:  6810 on 5 and 2510 DF,  p-value: < 2.2e-16
```

There is not much increase in this model Goodness of Fit compared to model with three lagged variables, it only account for 0.05% higher R-squared. The residual ACF plots also show similar result with the previous model.



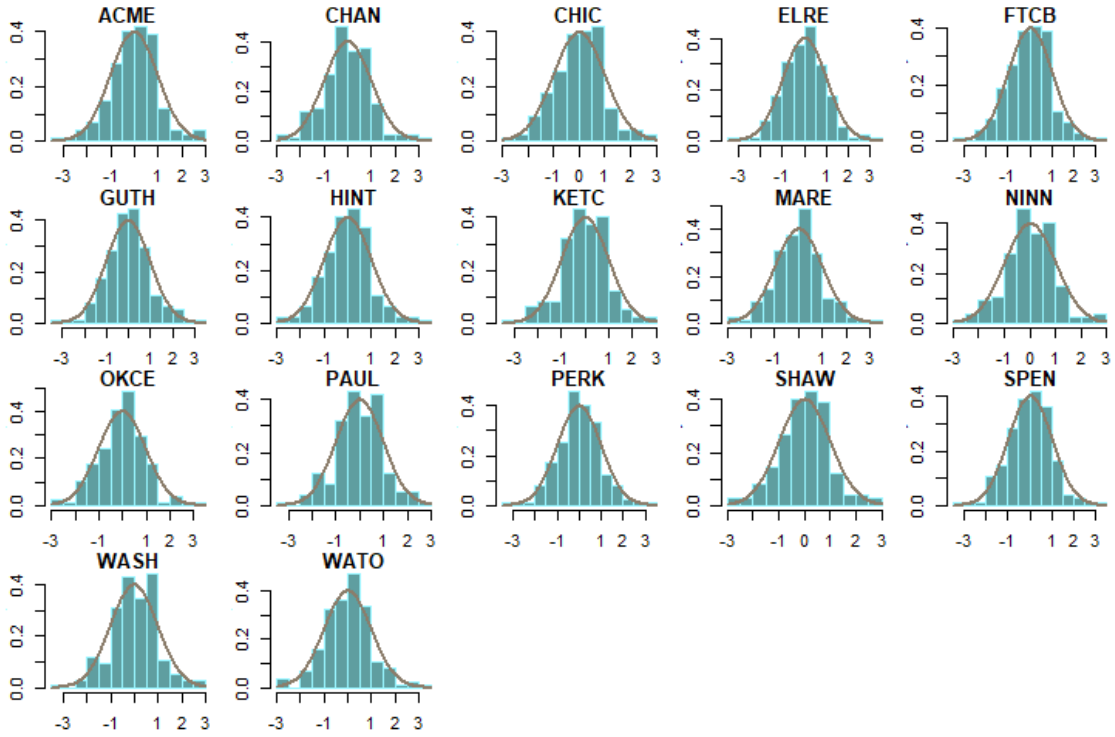
The following table is the summary of all previous models with regards to the number of lagged variables included in the model, the Goodness of Fit, and the independencies of residuals.

Lagged Variables in Model	Model Goodness of Fit		Durbin Watson Test		
	AIC	R-Squared	Autocorrelation	D-W Statistic	p-value
1 lagged variable	-5837.170	0.8984	0.3901934	1.219592	0
2 lagged variables	-6592.505	0.9258	0.1253156	1.747969	0
3 lagged variables	-6726.125	0.9308	0.00165793	1.995927	0.904
4 lagged variables	-6694.347	0.9313	-0.00644107	2.012828	0.766

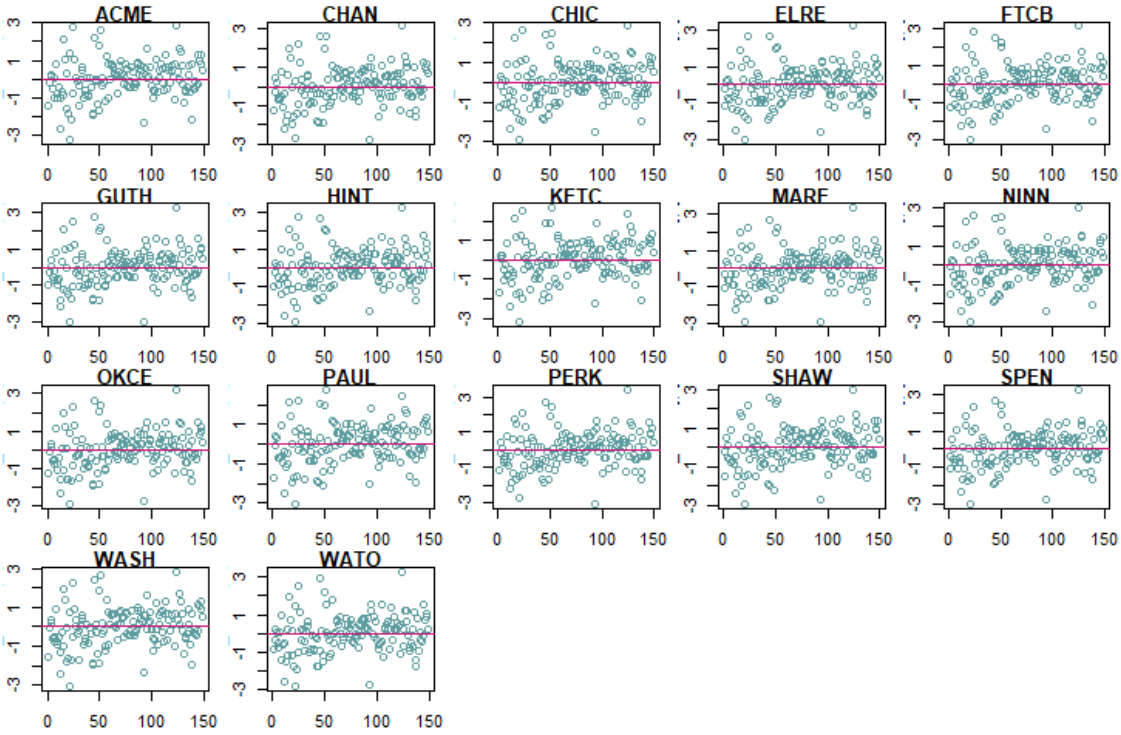
The summary shows that model with three lagged variables and elevation as the covariates returns the best Goodness of Fit according to AIC and more independent residuals according from Durbin Watson Test. Thus, this model is appropriate for the data and it is chosen as the final model.

2. Residuals Exploration

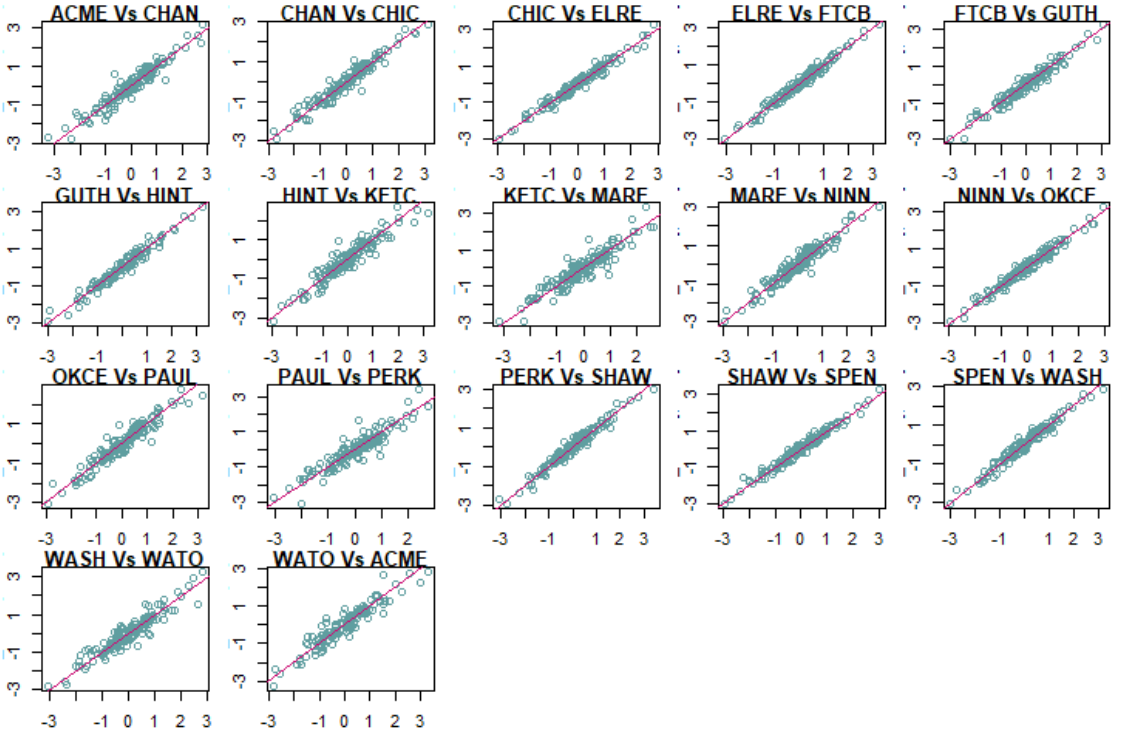
After getting the best model which has three lagged variables and elevation as the covariates, further analysis can be done throughout the residuals. The model is a combination of 17 stations, thus, the standardized residuals can be computed according to the mean and standard deviation of residuals from each station. Thus, there exist 17 series of standardized residuals which each represents its corresponding station. These 17-standardized residuals are assumed follow multivariate distributions. The following graphs show the histogram of each standardized residuals.



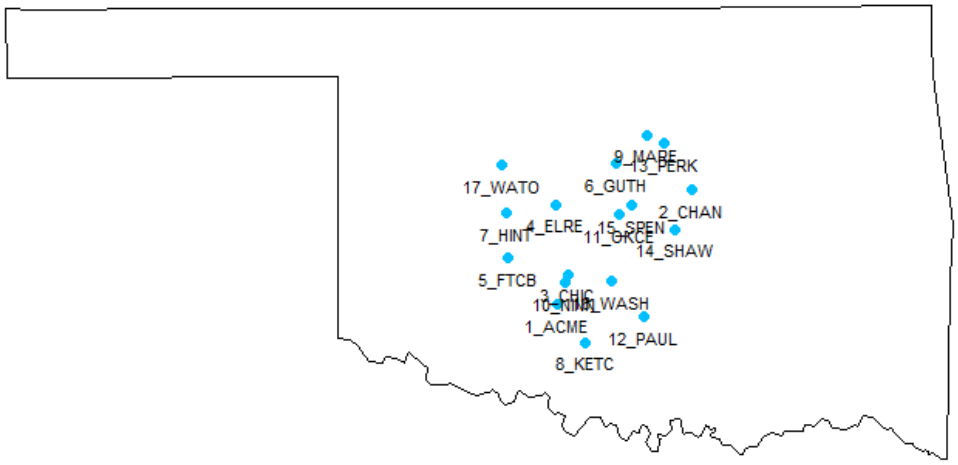
Looking at the distribution of each standardized residuals, each has mean close to zero and variance close to one. It shows that the standardized residuals are independent within a station. As also shown in the following scatter plots.



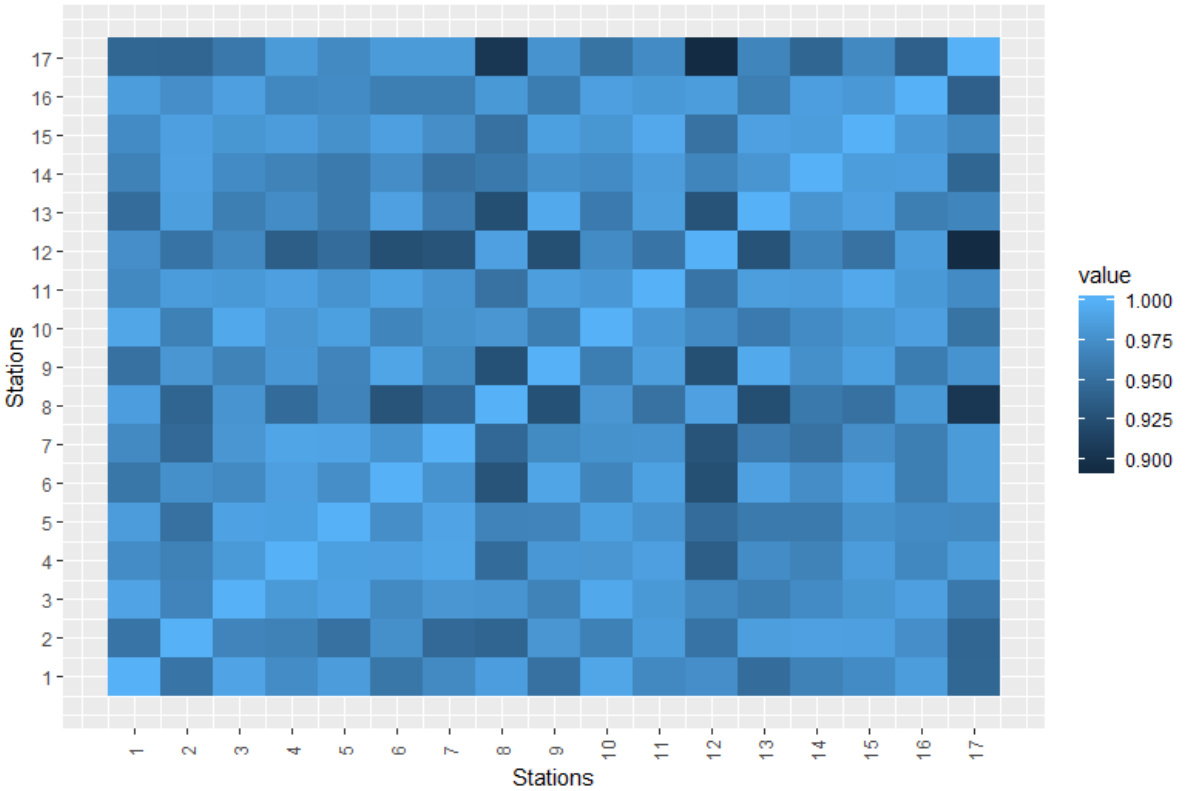
However, the following scatter plots of standardized residuals of two stations show that there exists a trend between two series of standardized residuals from different stations. This show that the standardized residuals of a station not independent with the other stations.



The following map show the location of each station in Oklahoma State, USA. Since there exist spatial dependencies in the model, it is assumed that the covariance of the standardized residuals between two stations is determined from the spatial effects, such as the distance between each station.



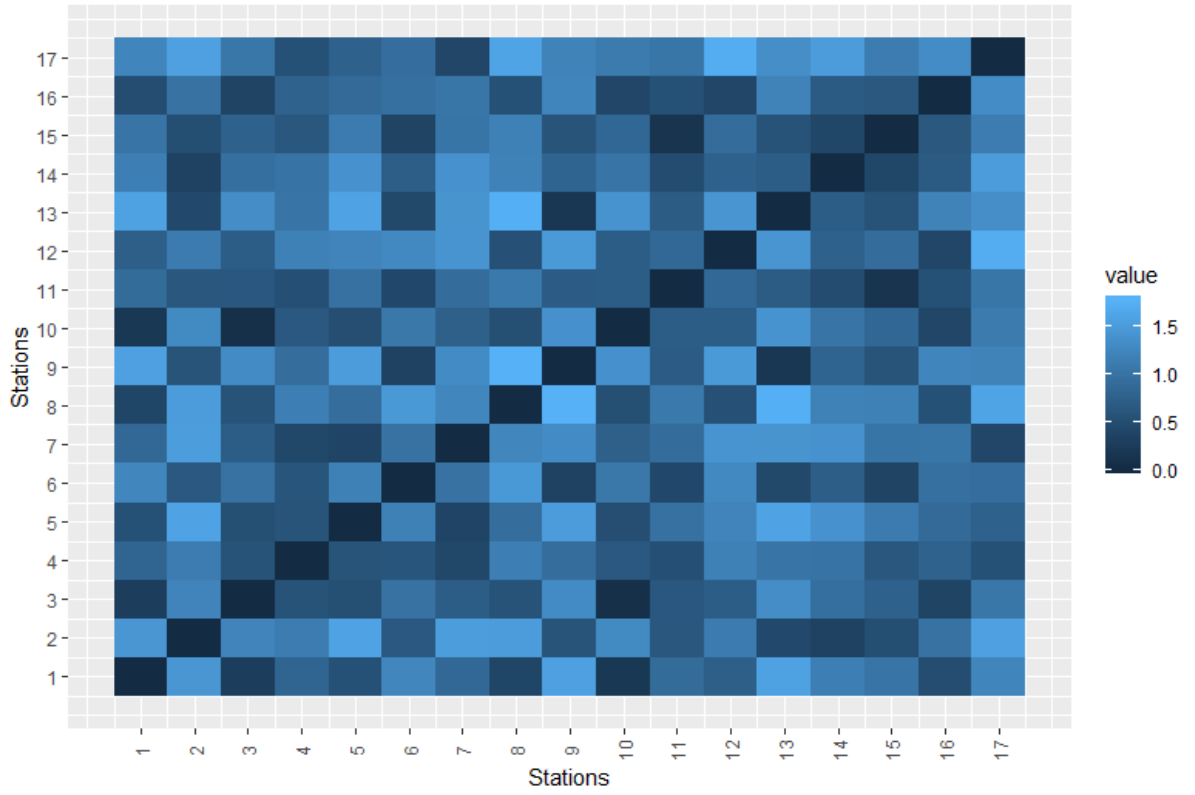
The correlation of standardized residuals for each station can be seen in the following diagram. For stations that are closely related according to spatial distance tend to have high correlation of standardized residuals.



All the above standardized residuals exploratory analysis strengthens the assumption that the standardized residuals for each station follow multivariate normal distribution with mean zero and covariance matrix $\sum$.

3. Parameters Estimate for Residuals Covariance Matrix

As explained in the methodology and evaluated in the previous analysis that the covariance matrix of standardized residuals follows a distribution based on the distance between stations. There are two parameters involved r and α and a distance matrix. The following diagram visualize the normalized distance matrix (1/100000*distance).



Parameter r and α then obtained using maximum likelihood approach. It starts from compute the covariance matrix from initial values of both parameters, then find the new parameters that minimize the log-likelihood function for multivariate normal density of the standardized residuals with mean 0 and the initial covariance matrix. Then back to count the new covariance matrix with the new parameters. These processes repeated iteratively until the value of both parameters converged. The following result shows the estimate of r and α that are converged after 18 iterations. Based on this approach it is obtained that $\hat{r}$ = 0.02787132 and $\hat{\alpha}$ = 0.85730123.

```
$`minimum`
[1] -1287.811

$estimate
[1] 0.02787132 0.85730123

$gradient
[1] -0.0035505377  0.0001657997

$code
[1] 3

$iterations
[1] 18
```

4. Confidence Intervals for Parameters Estimate

Since there always be errors in parameter estimation, then the next step is to construct confidence interval for the parameters of covariance matrix, r and α. The confidence interval can be obtained through bootstrap approach. The main sample is the standardized residuals of all stations. Bootstrap samples are drawn from this matrix as many as 1999 times and from each there are corresponding bootstrap parameter estimates for r and α. The confidence intervals then are constructed from 1999 pairs of bootstrap parameter estimates of r and α. The following result shows the bias and standard error of each parameter.

```
ORDINARY NONPARAMETRIC BOOTSTRAP


Call:
boot(data = stdres_station, statistic = parameters, R = 1999)


Bootstrap Statistics :
      original        bias    std. error
t1* 0.02787132  0.0008776141 0.006380696
t2* 0.85730123 -0.0035867265 0.044947926
```

There are four types of 95% confidence intervals resulted from bootstrap approach as shown in the following result.

Confidence interval for r

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates

CALL :
boot.ci(boot.out = bootstrap, index = 1)

Intervals :
Level      Normal              Basic
95%   ( 0.0145,  0.0395 )   ( 0.0216,  0.0317 )

Level     Percentile            BCa
95%   ( 0.0241,  0.0341 )   ( 0.0242,  0.0356 )
Calculations and Intervals on Original Scale
```

Confidence interval for α

```
BOOTSTRAP CONFIDENCE INTERVAL CALCULATIONS
Based on 1999 bootstrap replicates

CALL :
boot.ci(boot.out = bootstrap, index = 2)

Intervals :
Level      Normal              Basic
95%   ( 0.7728,  0.9490 )   ( 0.7722,  0.9465 )

Level     Percentile            BCa
95%   ( 0.7681,  0.9424 )   ( 0.7769,  0.9516 )
Calculations and Intervals on Original Scale
```

The following table shows the summary of the four types of confidence intervals for both parameters of residuals covariance matrix, r and α.

Confidence Intervals Types	Parameter			
	r		α	
Normal approximation	0.0145	0.0395	0.7728	0.9490
Basic (or residual)	0.0216	0.0317	0.7722	0.9465
Percentile	0.0241	0.0341	0.7681	0.9424
Bias corrected and accelerated percentile method (BCa)	0.0242	0.0356	0.7769	0.9516

The differences of confidence intervals between each type have no significant difference one to another. However, the confidence intervals as returned from percentile and basic have the smallest length and the one from normal approximation are the longest of all.

R Syntax

1. Upload dataset

```
weather = read.csv(file="./weather_data.csv")
stations = read.csv(file="./geoinfo_stations.csv")
```

2. Model Approach 1

```
intercept = 1:17
beta1 = 1:17
for (i in 1:17) {
  col = i + 3
  y = weather[2:152,col]
  ylag1 = weather[1:151,col]
  lm = lm(y ~ ylag1)
  intercept[i] = lm$coefficients[1]
  beta1[i] = lm$coefficients[2]
}

alpha = intercept
lat = stations[,8]
lon = stations[,9]
elev = stations[,10]
lm_alpha = lm(alpha ~ lat + lon + elev )
summary(lm_alpha)
pred_int = lm_alpha$fitted.values

mean_stdres = 1:17
sd_stdres = 1:17
pred = matrix(nrow = 151, ncol = 17)
res = matrix(nrow = 151, ncol = 17)
stdres = matrix(nrow = 151, ncol = 17)
for (i in 1:17) {
  col = i + 3
  y = weather[2:152,col]
  ylag1 = weather[1:151,col]
  pred[,i] = pred_int[i] + beta1[i]*ylag1
  res[,i] = y - pred[,i]
  mean_res = rep(mean(res[,i]),151)
  sd_res = sd(res_weat[,i])
  stdres[,i] = (1/sd_res)*(res[,i] - mean_res)
  mean_stdres[i] = mean(stdres[,i])
  sd_stdres[i] = sd(stdres[,i])
}
```

3. R-squared Model Approach 1

```
n = 151*17
act = 1:n
predict = 1:n
r1 = 1
r2 = 151
for (i in 1:17) {
  col = i + 3
  act[r1:r2] = weather[2:152,col]
  predict[r1:r2] = pred[,i]
  r1 = r1 + 151
  r2 = r2 + 151
}

actual = act
```

```
preds <- predict
rss <- sum((preds - actual) ^ 2)  ## residual sum of squares
tss <- sum((actual - mean(actual)) ^ 2)  ## total sum of squares
rsq <- 1 - rss/tss
rsq

rsq_app1 = 1:17
for (i in 1:17) {
  col = i + 3
  actual = weather[2:152,col]
  preds <- pred[,i]
  rss <- sum((preds - actual) ^ 2)  ## residual sum of squares
  tss <- sum((actual - mean(actual)) ^ 2)  ## total sum of squares
  rsq_app1[i] <- 1 - rss/tss
}
```

4. Model Approach 2

```
n = 151*17
yy = 1:n
yylag1 = 1:n
LAT = 1:n
LON = 1:n
ELEV = 1:n

r1 = 1
r2 = 151
for (i in 1:17) {
  col = i + 3
  yy[r1:r2] = weather[2:152,col]
  yylag1[r1:r2] = weather[1:151,col]
  LAT[r1:r2] = c(rep(stations[i,8],151))
  LON[r1:r2] = c(rep(stations[i,9],151))
  ELEV[r1:r2] = c(rep(stations[i,10],151))
  r1 = r1 + 151
  r2 = r2 + 151
}
lm_2 = lm(yy ~ yylag1 + LAT + LON + ELEV)
res = lm_2$residuals
stdres = stdres(lm_2)
summary(lm_2)
```

5. R-squared Model Approach 2

```
pred <- lm_2$fitted.values
prediction = matrix(pred, nrow = 151, ncol = 17)

rsq_app2 = 1:17
for (i in 1:17) {
  col = i + 3
  actual = weather[2:152,col]
  preds <- prediction[,i]
  rss <- sum((preds - actual) ^ 2)  ## residual sum of squares
  tss <- sum((actual - mean(actual)) ^ 2)  ## total sum of squares
  rsq_app2[i] <- 1 - rss/tss
}
```

6. Model with 2 lagged variables and forward stepwise

```
n = 150*17
yy = 1:n
yylag1 = 1:n
yylag2 = 1:n
LAT = 1:n
LON = 1:n
ELEV = 1:n

r1 = 1
r2 = 150
for (i in 1:17) {
  col = i + 3
  yy[r1:r2] = weather[3:152,col]
  yylag1[r1:r2] = weather[2:151,col]
  yylag2[r1:r2] = weather[1:150,col]
  LAT[r1:r2] = c(rep(stations[i,8],150))
  LON[r1:r2] = c(rep(stations[i,9],150))
  ELEV[r1:r2] = c(rep(stations[i,10],150))
  r1 = r1 + 150
  r2 = r2 + 150
}


lm_3 = lm(yy ~ yylag1 + yylag2 + LAT + LON + ELEV)
res = lm_3$residuals
stdres = stdres(lm_3)
summary(lm_3)

lm_step = stepAIC(lm_3, direction = "forward")
summary(lm_step)
```

7. Model with 3 lagged variables

```
n = 149*17
yy = 1:n
yylag1 = 1:n
yylag2 = 1:n
yylag3 = 1:n
LAT = 1:n
LON = 1:n
ELEV = 1:n

r1 = 1
r2 = 149
for (i in 1:17) {
  col = i + 3
  yy[r1:r2] = weather[4:152,col]
  yylag1[r1:r2] = weather[3:151,col]
  yylag2[r1:r2] = weather[2:150,col]
  yylag3[r1:r2] = weather[1:149,col]
  LAT[r1:r2] = c(rep(stations[i,8],149))
  LON[r1:r2] = c(rep(stations[i,9],149))
  ELEV[r1:r2] = c(rep(stations[i,10],149))
  r1 = r1 + 149
  r2 = r2 + 149
}

lm_4 = lm(yy ~ yylag1 + yylag2 + yylag3 + ELEV)
res4 = lm_4$residuals
summary(lm_4)
```

8. ACF Plots

```
par(mfrow = c(4, 5), oma = c(2,2,2,2), mar = c(2,2,1,1), mgp = c(2,1,0))
station_name = colnames(weather[,4:20])
res4 = lm_4$residuals
res4_mat = matrix(res4, ncol = 17, nrow = 149)
for (i in 1:17) {
  acf(res4_mat[,i])
  title(main = station_name[i])
}
```

9. Histogram Plots

```
for (i in 1:17) {
  hist(stdres_station[,i], breaks = 10, col = "cadetblue", border = "cadetblue1", prob
= TRUE, main = station_name[i])
  curve(dnorm,col="bisque4", lwd=2, add=TRUE)
}
```

10. Scatter Plots Residuals

```
for (i in 1:16) {
  plot(stdres_station[,i], stdres_station[,i+1], col = "cadetblue")
  titl = paste(station_name[i], station_name[i+1], sep = " Vs ")
  title(main = titl)
  abline(0,1, col = "deeppink3")
}
plot(stdres_station[,17], stdres_station[,1], col = "cadetblue")
titl = paste(station_name[17], station_name[1], sep = " Vs ")
title(main = titl)
abline(0,1, col = "deeppink3")
```

11. Maps

```
par(mfrow = c(1, 1))
sta_name = stations$stid
label = 1:17
for (i in 1:17) {
  label[i] = paste(label[i], sta_name[i], sep="_")
}
map_1 = map('state', region = 'Oklahoma')
with(stations, points(nlat ~ elon, pch = 19, col = 'deepskyblue'))
text(stations$elon, y = stations$nlat, label, pos = 1, font=1, cex = 0.7 )
```

12. Correlation Matrix

```
cov = cov(stdres_station[,1:17])
melted_cov <- melt(cov)
# View(melted_cov)
ID = 1:17
ggplot(data = melted_cov, aes(x=Var1, y=Var2, fill=value)) +  geom_tile() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  scale_x_continuous("Stations", labels = as.character(ID), breaks = ID) +
  scale_y_continuous("Stations", labels = as.character(ID), breaks = ID)
```

13. Distance Matrix

```
for (row in 1:17) {
  for (col in 1:17) {
    stasion_1 = c(stations[row,9], stations[row,8])
```

```
    stasion_2 = c(stations[col,9], stations[col,8])
    distmat[row,col] = distVincentyEllipsoid(stasion_1, stasion_2)
  }
}
adj_distmat = 1/100000*distmat

melted_dist <- melt(adj_distmat)
ID = 1:17
ggplot(data = melted_dist, aes(x=Var1, y=Var2, fill=value)) +  geom_tile() +
  theme(axis.text.x = element_text(angle = 90, vjust = 0.5)) +
  scale_x_continuous("Stations", labels = as.character(ID), breaks = ID) +
  scale_y_continuous("Stations", labels = as.character(ID), breaks = ID)
```

14. Parameter Estimates

```
fn <- function(theta) {
  sigma1 <- matrix(rep(0, 17*17), ncol = 17, nrow = 17)
  r = theta[1]
  a = theta[2]
  for (i in 1:17) {
    for (j in 1:17) {
      sigma1[i,j] = exp( -r * (adj_distmat[i,j]^a))
    }
  }
  if (a < 0 | a > 2) return(1000000)
  if (r < 0) return(1000000)
  loglik = sum(dmvnorm(stdres_station, mean = rep(0,17), sigma = sigma1, log = TRUE))
  return(-loglik)
}

param = nlm(fn, c(0.5, 0.5), hessian  = FALSE)
param$estimate[1]
param$estimate[2]
```

15. Bootstrap

```
parameters = function(x,i){
  bootsample = x[i,]
  fn <- function(theta) {
    sigma1 <- matrix(rep(0, 17*17), ncol = 17, nrow = 17)
    r = theta[1]
    a = theta[2]
    for (i in 1:17) {
      for (j in 1:17) {
        sigma1[i,j] = exp( -r * (adj_distmat[i,j]^a))
      }
    }
    if (a < 0 | a > 2) return(1000000)
    if (r < 0) return(1000000)
    loglik = sum(dmvnorm(bootsample, mean = rep(0,17), sigma = sigma1, log = TRUE))
    return(-loglik)
  }
  param = nlm(fn, c(0.5, 0.5), hessian  = FALSE)
  return(c(param$estimate[1], param$estimate[2]))
}

set.seed(123)
bootstrap = boot(data=stdres_station, statistic = parameters  ,R=1999)
boot.ci(bootstrap, index = 1)
boot.ci(bootstrap, index = 2)
```