Predicting temperature for Nebraska and Kansas

ROMAIN JUBAN

January 18, 2014

1 Exploring the data

The first step was to explore the data to understand it. The file training.csv contains time series of daily temperatures for 500 different locations. Dates range from January,1 1980 to December,31 2010. On Figure 1 we can see that the temperature varies seasonally with a frequency of 1 year (or 365 days) and that the mean temperature does not vary a lot (no particular trend).

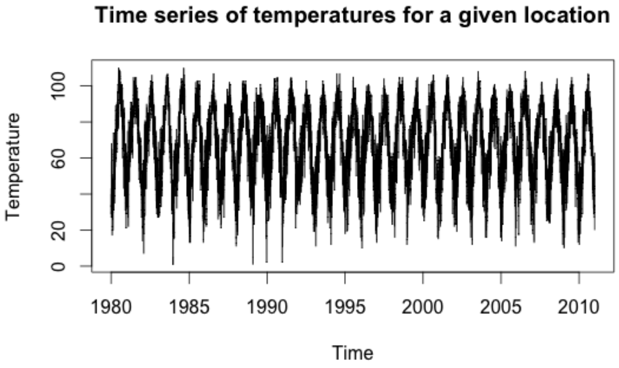

Figure 1: Time series of temperatures for a given location.

I decided to use time series in R and the package ”forecast”, which is powerful in handling time series and making predictions. Because the data has a periodicity of 1 year, we have to be careful with leap years. At the beginning, I set the frequency of time series at 365.25 to take them into account but this non-integer frequency caused many problems in counting the days. So I decided to get rid off all February 29’s, which is not significant for our prediction (we remove 9 points from more than 10000 observations and the prediction year (2011) is not a leap year). My training set has now 11315 rows and each year has 365 days.

Another interesting plot was a seasonal plot 2, which represents multiple seasonal time series for a given location on the same plot. We can note that the variance is quite large, with a spread that tends to be larger in winter (about 40 degrees) and smaller in summer (about 20 degrees).

From this plot, we cannot expect our predictions to be extremely accurate (¡10 degrees), given the uncertainty and variability of temperature for a given day over years.

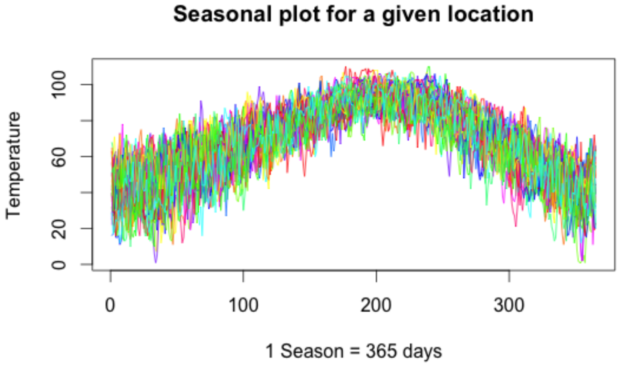

Figure 2: Seasonal plot for a given location.

2 Part I: Predicting temperature in 2011 for 500 different locations

After selection and validation of the model, I decided to use a STL model to make my predictions. This model performs seasonal trend decomposition, which was relevant for daily temperature time series (high seasonal period). STL decompositions are additive and use ARIMA process (Autoregressive integrated moving average).

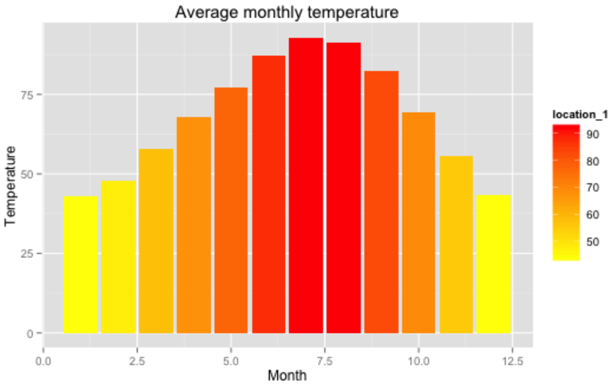

Figure 3: Monthly-averaged temperature for a given location.

As we can see on Figure 3, the month is a good feature to characterize the temperature. I added the month as a regressor variable in this model to smooth the fit.

Figure 4 represents the prediction using the STL method

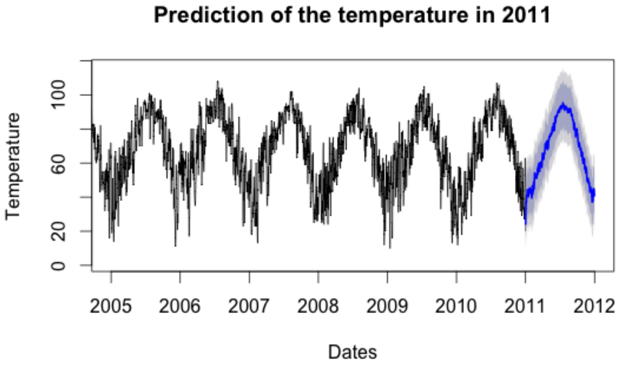

Figure 4: Prediction of the temperature in 2011.

for a given location. The shape of the curve is satisfactory and extreme values are comparable to the previous years.

After validating this model, I computed the temperature predictions for all locations for 2011.

3 Part II: Interpolating temperature for 50 new locations

To predict the temperature in 2011 for the new locations I have plotted on Figure 5 the sites to see how they are located and which interpolation method would be more appropriate.

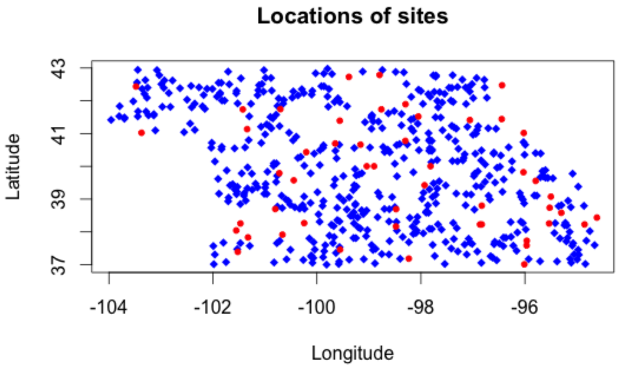

Figure 5: Spatial location of sites.

After selection and validation of the model, I performed the interpolation using Weighted K-Nearest Neighbors with K=2. This method uses the Minkowski distance.

4 Selection and Validation of the models used

To perform cross-validation of models, I decided to use a random sample of 10 locations to compute errors. To evaluate models I have used the Mean Squared Error (MSE).

4.1 Size of the training set used

At a first glance, it is not obvious whether the complete set (from 1980 to 2010) would lead to better predictions. A smaller dataset could better factor a short-term trend in temperature but could also lead to higher bias (underfitting). I decided to train the model on different sizes of training set (starting from 1980 - 2009 to starting from 2007 - 2009) and compute the test MSE for 2010.

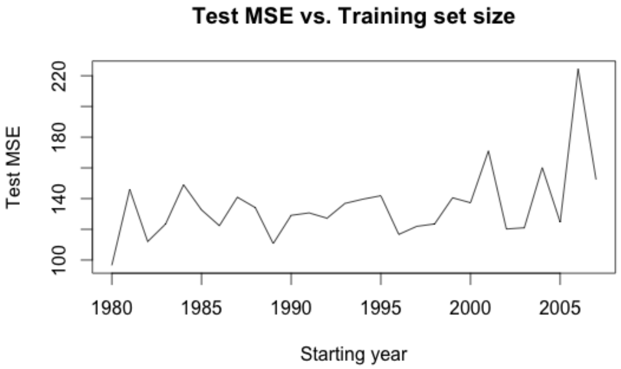

Figure 6: Test MSE vs. training set size.

We see on Figure 6 that the lowest MSE is obtained when we train our model on the full dataset.

4.2 Selection of the model for part I

I tried different methods usually used in predicting time series. Given the high seasonal period of the data, many models were not appropriate. For example ARIMA or ETS do not handle seasonality frequency of 365 days. I compared four different models: STL without month factored, STL with month factored, Linear Regression and Holt-Winters Exponential Smoothing. I trained those 4 models on 29 years (from 1980 to 2009) and computed the test MSE for 2010:

- STL without month factored: 96.4
- STL with month factored: 96.6
- Linear Regression: 100
- Holt-Winters Exponential Smoothing: 249

4.3 Selection of the best interpolating method for part II

I tried two different methods to interpolate results from the 500 sites to the 50 new locations.

Weighted K-Nearest Neighbors: It uses Minkowsky distance and Kernel functions to weight the neighbors according to their distances. Thus, it is possible to take into account the spatial distance among locations in the interpolation (this is not the case with a traditional KNN method). Here, the tuning parameter is the number of neighbors considered.

Spline interpolation with a GAM model: We fit a Generalized Additive Model using smoothing splines for the two components of position (longitude and latitude). The tuning parameter are the degrees of freedom of the splines.

I performed cross-validation using the 500 locations to tune the parameters and select the best model. I randomly chose 400 locations to train my two models and computed the test MSE on the remaining 100 locations.

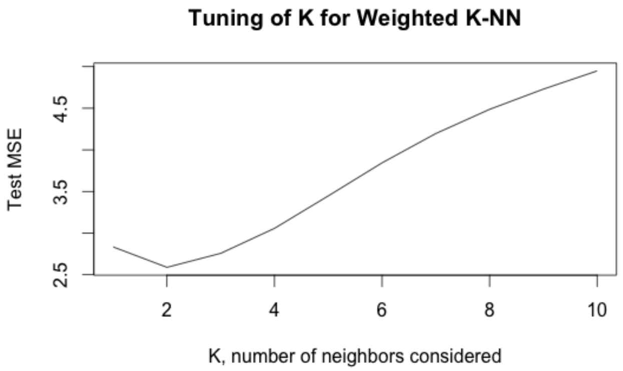

Figure 7: Tuning of the parameter K for Weighted K-NN.

The best model for Weighted K-NN is obtained with K=2 and the corresponding test MSE is 2.6.

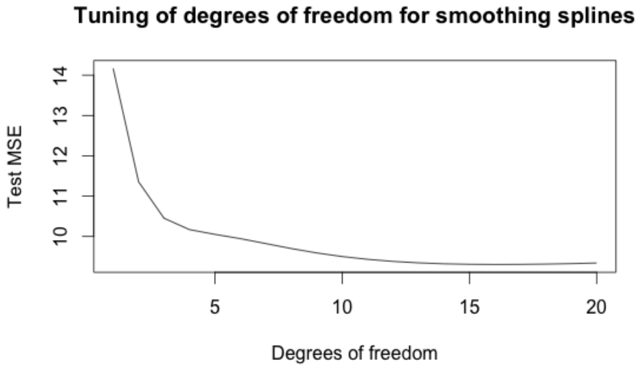

Figure 8: Tuning of degrees of freedom for smoothing splines.

The best model for smoothing splines is obtained with 16 degrees of freedom and the corresponding test MSE is 9.3.

5 Results analysis

To assess the accuracy of my results (values, magnitude of the seasonality) I have decided to plot my data spatially in 2011 (forecasted) and 2010 (observed).

Yearly-averaged temperature:

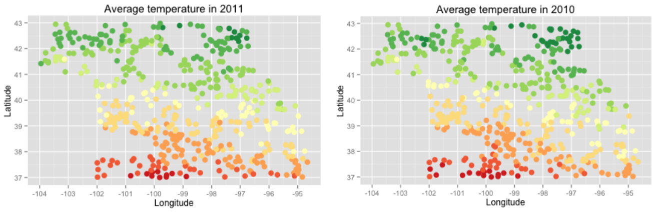

Figure 9: Yearly-averaged temperature in 2011 and 2010.

Minimum temperature:

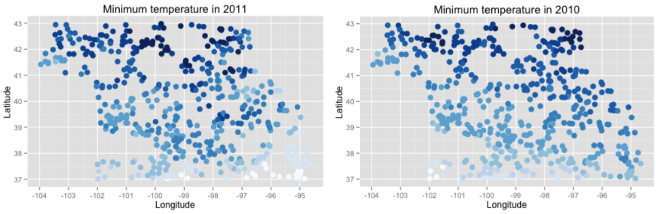

Figure 10: Minimum temperature in 2011 and 2010.

6 Possible improvements

It is difficult to improve the accuracy of the prediction one year ahead given only historical values and without more predictors. We saw that our best model made on average an error of less than 10 degrees Fahrenheit (or 5 degrees Celsius) for each daily temperature (test MSE smaller than 100), which is pretty good. A traditional weather forecast using Numerical Weather Prediction could make such an error one week ahead. Weather features like cloud cover, specific humidity, pressure would probably sharpen the predictions.

For the interpolation, the Weighted K-NN performs very well. We could try a Kriging Method but it might be very computationally intensive.

References

[1] *Package 'Forecast'*, `http://cran.r-project.org/web/packages/forecast/index.html`.

[2] *Package 'GAM'*, `http://cran.r-project.org/web/packages/gam/index.html`.

[3] *Package 'KKNN'*, `http://cran.r-project.org/web/packages/kknn/index.html`.

[4] *Package 'ggplot2'*, `http://cran.r-project.org/web/packages/ggplot2/index.html`.

[5] *Package 'RColorBrewer'*, `http://cran.r-project.org/web/packages/RColorBrewer/index.html`.