PROGRESS REPORT

DATA ANALYSIS: TRENDS IN WEATHER

MARCUS LORENZANA, ROBERT VASQUEZ

Project Description:

For our project we did a data analysis of weather data over a period of 14 years in the United States. Looking at different regions in the United States, we determined whether there are any interesting existing patterns and used Time Series Analysis to predict future patterns in temperature and precipitation. We also used linear regression to determine whether or not there is a positive or negative trend in temp/prec. By comparing other regions with each other, we were able to find difference in seasonal temperatures and precipitation, and make conclusions on the current and future data.

Data Set:

Our data, which contains monthly temperature means and total precipitation for the given month, was obtained from the National Climatic Data Center. The data spans 14 years and consists of numerous different weather stations in different areas of the United States. As a result, the original file is very populated (~1,000,000 records) and requires a lot of preprocessing in order to work properly in Weka.

Preprocessing:

For data cleaning, we first needed to delete records that had corrupted data. Many of the stations contained records for either temperature or precipitation with the value “-9999”. We found that stations which contained these values also contained the same errors for most of their months so the best course of action we believed was to remove the records entirely. The next step was to separate the data by state, given that there were over a million records in the main climate file. After separating the file into 50 files for each given state, the number of records for each file were typically around 10,000 with some less and some much more. Another Important step was to sort the data into ascending order by both month and year so that it is suitable for the Weka Time Series Analysis or Forecast tool. Although the data was in ascending order, since there are many stations each station has their own values for their given months. As a result, there are many duplicates for the exact same date and Weka will not accept that.

For the final step we merged these dates together by finding the mean of all the stations for each date. We essentially created buckets with each bucket containing a list of all records that match

that given date. We then iterated over those buckets and for each one, found the mean MNTM and TPCP for that month.

The detailed steps in preprocessing are as followed:

1. Run ClimateCleanStates to remove the -9999 values and distribute data to 50 files representing each state. Here, a HashMap is used with the State as the Key, and the File as the value. This makes it so that you don't have to close or flush the files after every row insertion and thereby speeding up the file processing significantly.
2. Run ClimateCleanDatesAscending and provide state abbreviation. This program will sort all of the data from all stations into complete ascending order (by date). For example, the first grouping of lines would be January, 2000. It will use the class ClimateRecord and ClimateComparator to sort the data in such a way.
3. Run ClimateCleanAverage and provide state abbreviation. This program will merge the duplicate date cells together by averaging all of them out and assigning it to that month. It will use the class ClimateRecord and ClimateRecordList in order to store the buckets, each containing a list of data for that month.

Time Series Analysis:

For our data analysis we used Time Series Analysis using Weka to find interesting patterns. Time Series Analysis is a process using multiple statistical and mathematical techniques in order to forecast future trends. With the Time Series Forecast tool in Weka, we can input how far into the future we want predicted data, and in our project we predicted up to two years or 24 units of time (24 months). Using the monthly temperature means and monthly rainfall means, we found trends for our data in the 14 years, as well as the future forecast.

Here is a sample of the graphs based off of different regions in the United States (West,Midwest,South,North-east). For West, we chose California, Texas for south, Iowa for Midwest, and New York for North-east.

Key:

Solid blue lines : Real precipitation data

Dashed blue lines : Predicted precipitation data

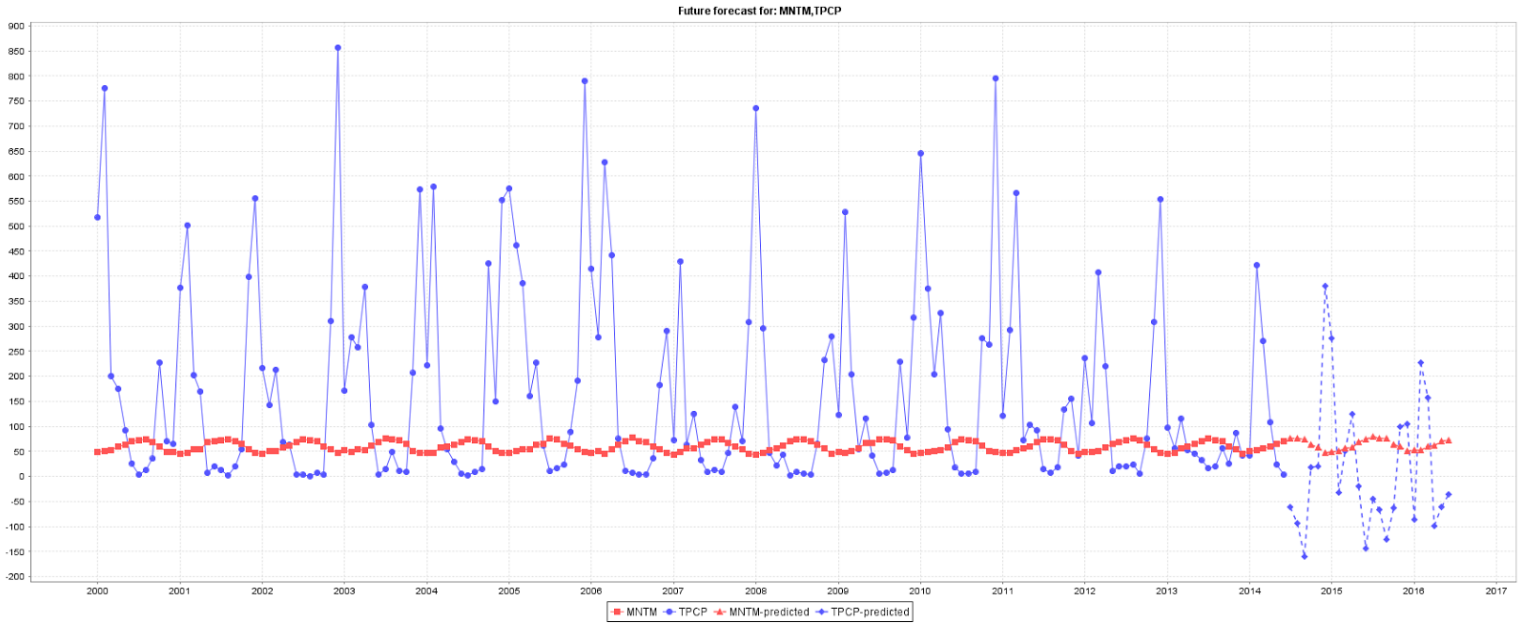
Solid red lines : Real temperature data

Dashed red lines : Predicted temperature data

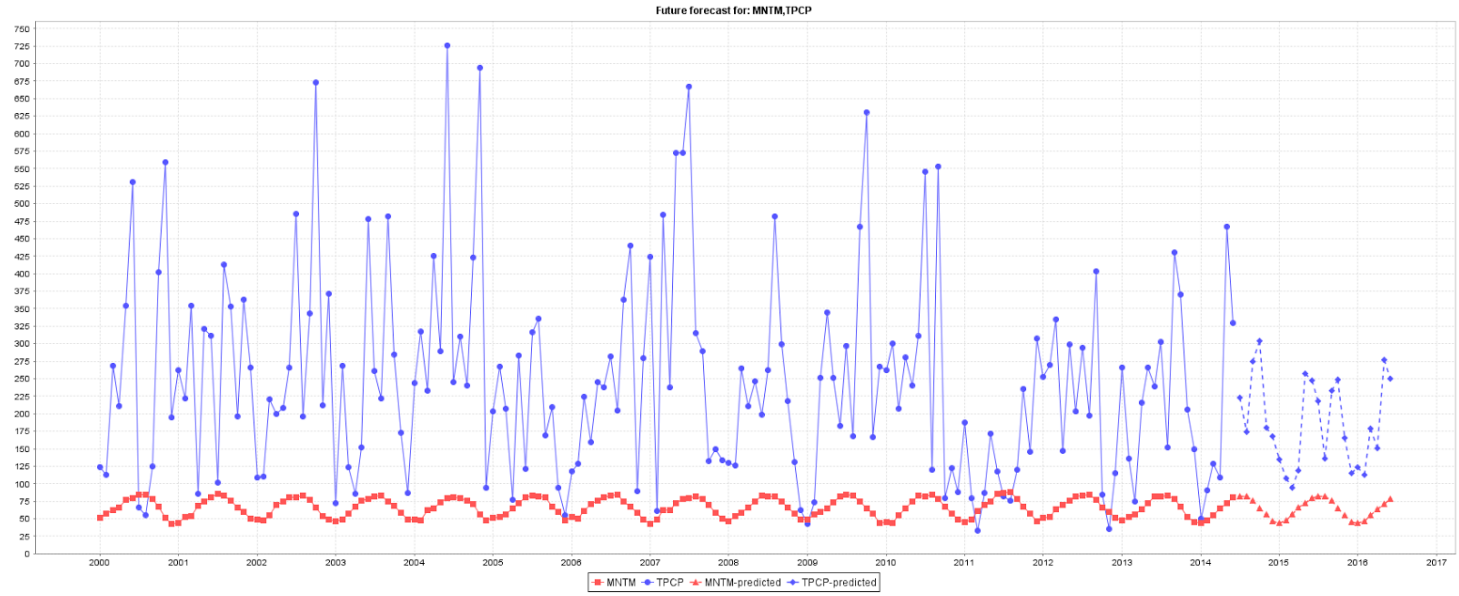
The detailed steps to reproduce the output below would be as followed:

1. Download Weka Development version. Open package manager and install Time Series Forecast near the bottom.
2. If would like to see how sorting of States works, run ClimateCleanStates
3. Run ClimateCleanDatesAscending and provide state abbreviation.
4. Run ClimateCleanAverage and provide state abbreviation
5. Open the new .arff file in Weka and use the Forecast tool. Set time units of forecast (optional).
6. Select attributes and click start.
7. Click Train future pred. tab to view the graph. Can also zoom in.

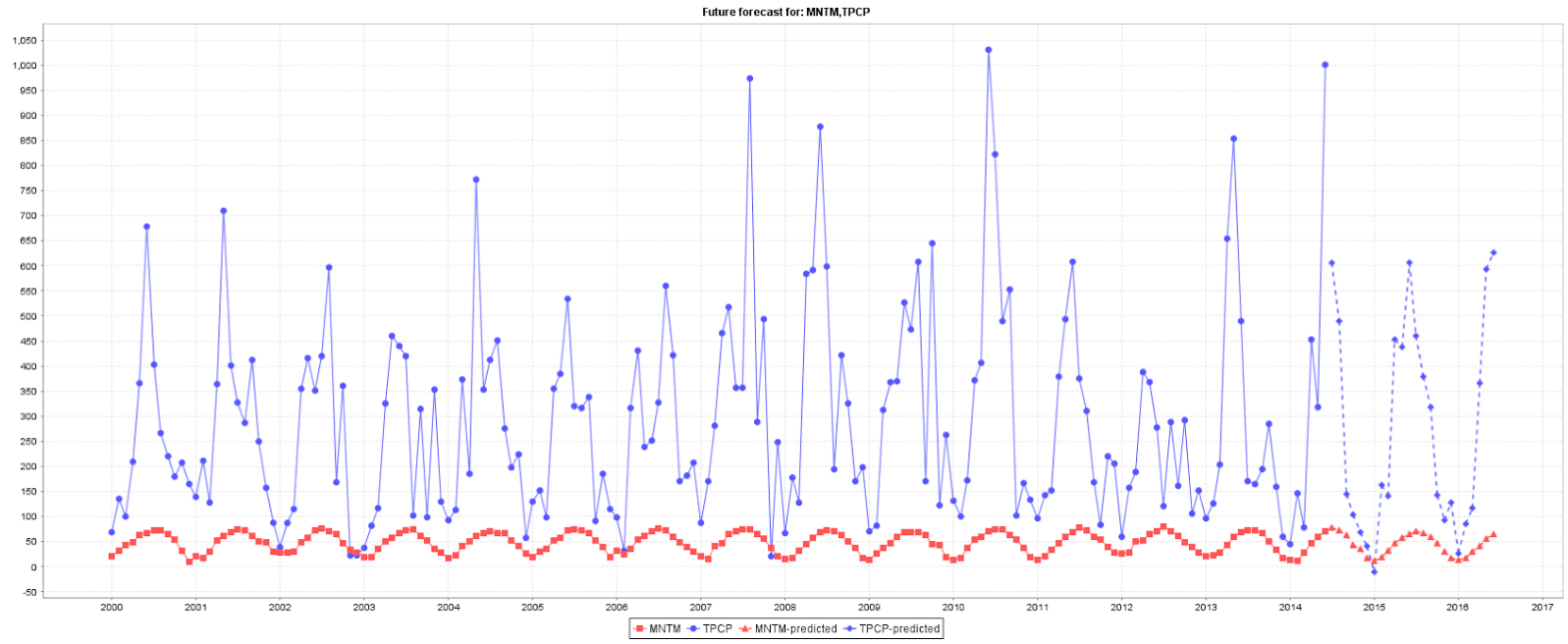
California:



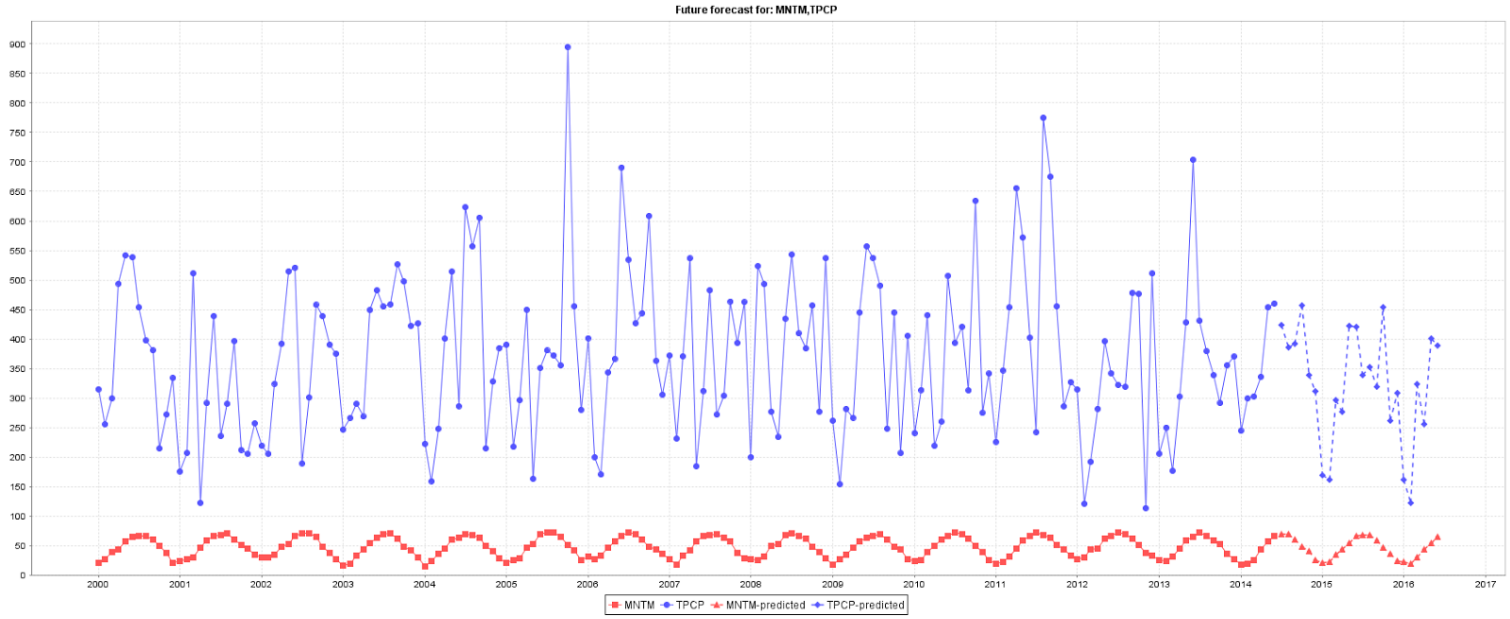
Texas:


Future forecast for: MNTM,TPCP
MNTM
TPCP
MNTM-predicted
TPCP-predicted

Iowa:


Future forecast for: MNTM,TPCP
MNTM
TPCP
MNTM-predicted
TPCP-predicted

New York:



Linear Regression:

We also decided to use linear regression to determine if there was any absolute negative or positive trends in temperature and precipitation. Our idea was that over time as a result of global warming, perhaps temperatures overall have increased over the past couple of years. We also applied the linear regression technique to precipitation in order to find an indicator of whether or not rainfall has increased or decreased, perhaps resulting in droughts in California, or increased frequency of storms in Florida.

Some sample data:

Correlation Coefficient TPCP California: 0.1083

Correlation Coefficient MNTM California: -0.183

Correlation Coefficient MNTM TPCP California: 0.6702

Correlation Coefficient TPCP MNTM California: 0.6541

Conclusion:

The bigger spikes in monthly rainfall, we found was usually due to major storms or hurricanes, which can be seen in states like Florida, a state prone to tropical storms. This also shows predictable hurricane seasons, which was also consistent with our data. For future weather in the next two years, the temperature remained consistent and fell into the predictable seasonal patterns. However, rainfall seemed to be slightly decreasing with less major spikes. Partially January of 2016 seems to be a very low year in rainfall for many states, with some predicted to have drought like conditions. Using linear regression, there was no hard evidence for TPCP and MNTM alone for each state although when running the linear regression with both attributes it does appear that there is a positive correlation as the values are closer to 1 than 0 (0.67). When looking at the graph, you can see that when the temperature falls for a period of time, there tends to be spikes in precipitation around that time too.

Sample Time Series Forecast values (* indicates predicted):

01-2013 44.38

02-2013 47.58

03-2013 55.23

04-2013 59.61

05-2013 64.49

06-2013 70.94

07-2013 75.46

08-2013 72.91

09-2013 69.6

10-2013 59.97

11-2013 53.98

12-2013 45.63

01-2014 51.49

02-2014 51.61

03-2014 55.67

04-2014 58.77

05-2014 65.41

06-2014 70.53

07-2014* 76.9899

08-2014* 75.4405

09-2014* 72.6816

10-2014* 63.5466

11-2014* 56.8063

12-2014* 47.9935

As you can see from the data above, the predicted values for 2014 seem to be fairly accurate and looks quite similar to 2013 values. When running precipitation future predictions seemed to somewhat follow the previous trends although it was much less accurate I believe because those large precipitation spikes don't always follow a set cycle.

Project Contributions:

Marcus Lorenzana:

- ClimateClean Project
    - ClimateAverage
    - ClimateCleanDates
    - ClimateCleanStates
    - ClimateDataComparator
    - ClimateRecord
    - ClimateRecordList
- Initial sample set of forecast models and linear regression
- Project Proposal
- Presentation Slides
- Final report
- Datasets

Robert Vasquez:

- The rest of the forecast models using ClimateClean Project
- Progress Report
- Final report

Please view the GitHub repository for complete commit history:

https://github.com/MeesterMarcus/DataMiningProject

Experience:

As far as experience goes, I enjoyed the freedom in doing a project of our choice. I also learned a lot about preprocessing and data mining in the process. I especially learned how to process files more effectively and got a little bit better at OOP. I do wish that we could have gone over Weka a bit more extensively as it was an important tool for the project. There also was very little contribution from Robert Vasquez unfortunately and I needed more help discovering patterns and applying data mining techniques as well as finding interesting patterns. Overall, the project did take a large portion of my time and having a more balanced workload on the project would have made the course more manageable for me.

Additional Information:

National Climatic Data Center Data Set:

https://www.ncdc.noaa.gov/cdo-web/datasets

Primary Time Series Analysis resource:

http://wiki.pentaho.com/display/DATAMINING/Time+Series+Analysis+and+Forecasting+with+Weka

Other Material for our project:

http://www.academia.edu/5173396/Effectiveness_of_Using_Data_Mining_for_Predicting_Climate_Change_in_Sri_Lanka

http://web.ornl.gov/sci/knowledgediscovery/ClimateDataMining/docs/Ganguly_ICDM-SSTDM-slides_2008.pdf