Chronology of a model building exercise - the 2008 AI Competition.

Matthew Pocernich *

National Center for Atmospheric Research, Boulder, CO, USA

* *Corresponding author address:* Matthew Pocernich, National Center for Atmospheric Research, Research Application Laboratory, Boulder, CO.

E-mail: pocernic@ucar.edu

ABSTRACT

This paper chronologically describes the steps taken in creating a predictive model for submission in the 2008 Artificial Intelligence - Prediction competition. The problem posed in the contest was to predict the type of precipitation (frozen, liquid or none) using information provided by radar. Primarily, randomForest models and variations of a neural network model were used. The final prediction was made using a combination of neural network models that attempted to exploit the assumption that the test data set was sampled from within the training data set.

1. Introduction

This paper chronologically describes the steps taken in creating a predictive model to create a forecast for submission in the 2008 Artificial Intelligence - Prediction competition. As the saying goes, getting there is half the fun. For this reason, mention will made of the of the many excursions, detours and road blocks that were encountered during this exercise. All models presented here were implemented using the R statistical programming language (R Development Core Team 2008). This open source and freely available language has become the *lingua franca* of the statistical research community. R operates on most operating systems including Mac OS, Linux and Windows. Most importantly, more than 1,500 packages have been created presenting a plethora of methods used in a variety of topics. The primary three packages used in this experiment are **nnet**, **randomForest**, and **verification**. The verification package was developed by people at the National Center for Atmospheric Research (NCAR) and consists of function useful in forecast verification. While it was developed within the context of weather model verification, many of the methods are commonly used in other fields such as biostatistics and engineering.

This paper is organized as follows. The problem and training data are discussed and summarized in Section 2. Results from the initial exploratory data analysis are presented in Section 3. Comments on the Pierce Skill Score (PSS) are given in Section 4. A brief overview of the statistical methods considered and used are discussed in 5. Closing comments and discussion are presented in Section 6. The methods used here are only briefly. References are provided for more detailed information about the statistical methods.

2. Problem statement and dataset.

The goal of the contest is to correctly predict the type of precipitation occurring bases on radar based measurements. The precipitation was categorized as being either frozen, liquid or none. Information was not provided about how the observations were made or collected. Potential predictors include the following.

- 2 m air temperature (in degrees Celcius)
- Relative humidity (in percent)
- u (East-West) component of the wind (West wind is positive)
- v (North-South) component of the wind (South wind is positive)
- freezing level (in m MSL)
- tilt, radar elevation angle (in degrees)
- range from radar (in km)
- Azinuth from radar (in degrees, 0 degrees os North)
- Height of radar data (in km)
- differential reflectivity (in dB)
- RhoHV, cross-crrelation coefficient (unitless, may exceed 1 due to radar calibration errors)
- Kdp, specific differential phase (degrees per km)

- Z, reflectivity (in dBZ)

- observed, observed precipitation type (in training data only).

Freezing, liquid and no precipitation was observed approximately 58%, 28% and 13% respectively.

3. Exploratory Data Analysis

One of the guiding rules in statistics is to know your data. In the real world, this type of project would begin with discussions with the scientists who have collected data (or ideally who are about to collect the data). In the contest setting, a forensic approach is used to uncover the characteristics of the data. Much of this effort is aimed at discovering patterns, relations between variables as well as idiosyncrasies with the data.

a. Spatial Aspects

With latitude and longitude provided with each record, plotting the data points on a map provided a quick summary of the distribution of the data (Figure 1). Initially, I assumed that the observations were aircraft based and the high density of datapoints was centered at an airport. With the help of Google Earth and information from contest sponsors, I found this not to be the case. What I had thought to be approach routes - turned out to be roads. Characteristically, the observations were not found to be evenly dispersed. This is very common in meteorological observations. Observations which are made by people tend to be clustered around cities. Observations made by aircraft tend to be clustered around

airports and major routes.

The importance of spatial clusters of observations is that these observations may be describing the same feature of event. This may result in the over-representation of these events. Statistically, this addresses the issue of independent vs. dependent data. There may also be spatial differences in the meteorological process. The climatology of an event may vary. This information may be useful in creating a baseline forecast.

While Figure 1 reveals that the observations were not evenly dispersed - this information was not incorporated into the model or model building process. An idea that was explored - but not implemented was the stratification of the data into independent areas. Data randomly selected from each area might provide a less bias spatial sample of the data.

b. Relationship between data

To study the relation between variables and between variables and the observations, a pairs plot (sometimes referred to as a matrix plot) was created. This type of plot is best explained through illustration. A pairs plot of the first six variables is shown in Figure 2. Variable names are listed along the diagonal grid boxes. Across the rows, these variable are displayed along the y-axis. Along the columns, these variables are summarized on the x-axis. The plots presented in the upper right portion of the plot are reflected in the lower left corner of the plot. The observations are displayed by using color to code the points on the graphs.

This pairs plot reveals some interesting features. By looking at the column with freezing level (frzl), one sees some interesting structure in the data. Freezing level appear to occur

at finite levels. Within these levels, precipitation type seems to be somewhat uniform. For example, at an elevation of 3000, most observation are for frozen. Not surprisingly, the tilt and the range of the radar are directly related.

Within these relations, there does not appear a good, simple solution. Ideally, one might find a pair of variables that would stratify the observation types into distinct groupings.

c. Time series

While the data was not described as being presented in any particular order, plotting the data in series can be informative. Figure 3 contains plots of temperature (tmpc) and a component of the wind speed (urel) for both the training data and the testing data. There are some striking features about these plots. First, there are series of sequences in which values remain fixed at a single value. One also notices that between variables, these period of constant values are the same. This high degree of correlations between variables and within a sequence suggest that the data was not randomly sampled. This is not surprising. Much data that is collected in field experiments by design has serious biases associated with it. For example, data reported by pilots is biases towards more extreme events and is biased towards locations around major airports.

These periods of extreme agreement suggested to me that the data may have been gathered on a small number of time periods. A dummy variable was created by multiplying the wind speed components, the temperature and the relative humidity variables. Even considering the resolution of the different instruments, it seem unlikely that if the events were the cases independent this product would ever be exactly identical. What was found was that

the instead of 847 independent records, there are approximately 135 "events". Moreover the most represented event contained 80 records and the 25 largest events included 636 records.

In retrospect, this is not surprising. I am guessing that the temperature wind and relative humidity fields were taken a single location. Perhaps the observations were aggregated within a single hour - so it might be reasonable to expect that multiple observations would be associated with identical with a single weather station.

Looking at the test data as plotted sequentially in Figure 3, one also notices that the test data seems to be much more dispersed than the training data. More will be said about this later.

4. Pierce Skill Score

The effectiveness of the models was judged by the Pierce Skill Score (PSS). The PSS can be defined as the difference between the hit rate and the false alarm rate. The hit rate expresses the frequency in which the model is correctly forecasts an event to occur when in fact it does occur. A major issue with the PSS is that it does not reward a forecast of a rare event. For example, if an event such as tornadoes occur 1% of the time, correctly forecasting such an event is considerably more difficult than forecasting an event not to occur. The PSS does not differentiate between such correct forecasts.

Aspects of the PSS score were studied for the following reason. Observations of "none" occurred only 13% of time. Models such as randomForests as well as other clustering methods produce a probability of a record belonging in each category. The single forecast of liquid, frozen or none is given to the outcome with the highest probability. The idea was explored

of requiring greater certainty in forecasting the more unlikely events since these events are less likely to occur. This can correctly be viewed as a way of gaming the PSS. This was not successfully incorporated into the model.

5. Statistical models

Two major types of models were used in this study. Random Forests (Breiman 2002) create a "forest" (or ensemble) of regression trees from a randomly selected subset of variables. A forecast is created by aggregating the votes from all regression trees. Parameters that can be varied in this model include strata and weights.

Much effort was given to adjusting the **strata** and **sampsize** parameters. **strata** is a list of factors which guide stratified sampling. With stratified sampling, **sampsize** dictates how many samples would be take from each group. This was done to encourage a diversity of events be used in the creating the model. Ideally, this would create a more robust model. Unfortunately, this did not result in any notable improvements in the PSS.

A multinomial log-linear model (**multinom**) was used to accommodate the categorical data in the observations and categorical factors - the event ids, in the predictors. Neural networks are used to fit this model (Ripley 1996; Venables and Ripley 2002). The most desirable attribute of this model is that a large number of factors - such as event ids can be used. Fitting such a large number of variables is typically a disadvantage

With multinomial log-linear model, the effects of varying the **weights** parameter was explored. This parameter describes the weight placed on each record. Efforts were made to reduce the importance given to data collected during a very large events and increase the

weight given to data collected on events with fewer observations. Varying the weights did not appear to have any great effect on the performance of the model.

All of the data including the event id was used to create a multinomial model. Typically, The forecast made with this model performed very well with a PSS value of 0.52. Next the data was randomly partitioned into 5 groups. The model was build on 4/5 of the data and a forecast was made on the 1/5 of the data that was not used in the model selection. The PSS value fell to values of 0.3. This illustrates the dangers of an overfit model.

6. Results and Comments

After much deliberation, the following type of hybrid model was used to create the submitted forecast. With the training data, two multinomial models were created. One used the event id as a variable, the second did not. Of the 363, records in the testing data, 334 matched with an existing event id from the training dataset. These could be used with the first model. The remaining cases were used in the second model which did not incorporate event ids.

The decision to use this type of model was based on the following rationale. From Figure 3 and from comparing the values of the event ids of the training and test data, it appeared that the testing data was sampled from within the time period of the training data. The testing data was not new data such as might be available if the models were build on 3 years of data then tested on the subsequent year of data. For this reason, I assumed that perhaps using an overfit model might have some benefits. The model wouldn't be useful in forecasting most new conditions but might be good at forecasting conditions very similar to

those which had happened. This is not a realistic model for operational use. The likelihood of conditions in the future having identical event ids is not likely. This was a model created to win this contest.

Additionally, after submitting an official set of forecasts a second forecast was made using a randomForests model with no modifications or adjustments of any kind made to the randomForest function. Ironically, this was the first model that was tried and possibly the simplest. For all the time spent studying the effects of different weighting schemes, different strata and in general scheming, little improvement was made.

REFERENCES

Breiman, L., 2002: Random forests. *Machine Learning*, **45**, 5–32.

R Development Core Team, 2008: *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing, URL `http://www.R-project.org`, ISBN 3-900051-07-0.

Ripley, B., 1996: *Pattern Recognition and Neural Networks*. Springer.

Venables, W. and B. Ripley, 2002: *Modern Applied Statistics with S*. Springer.

List of Figures

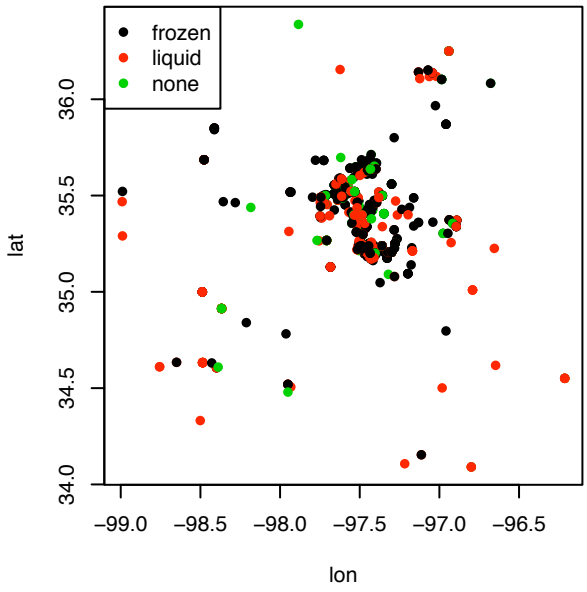

FIG. 1. Map indicating the location of observations.

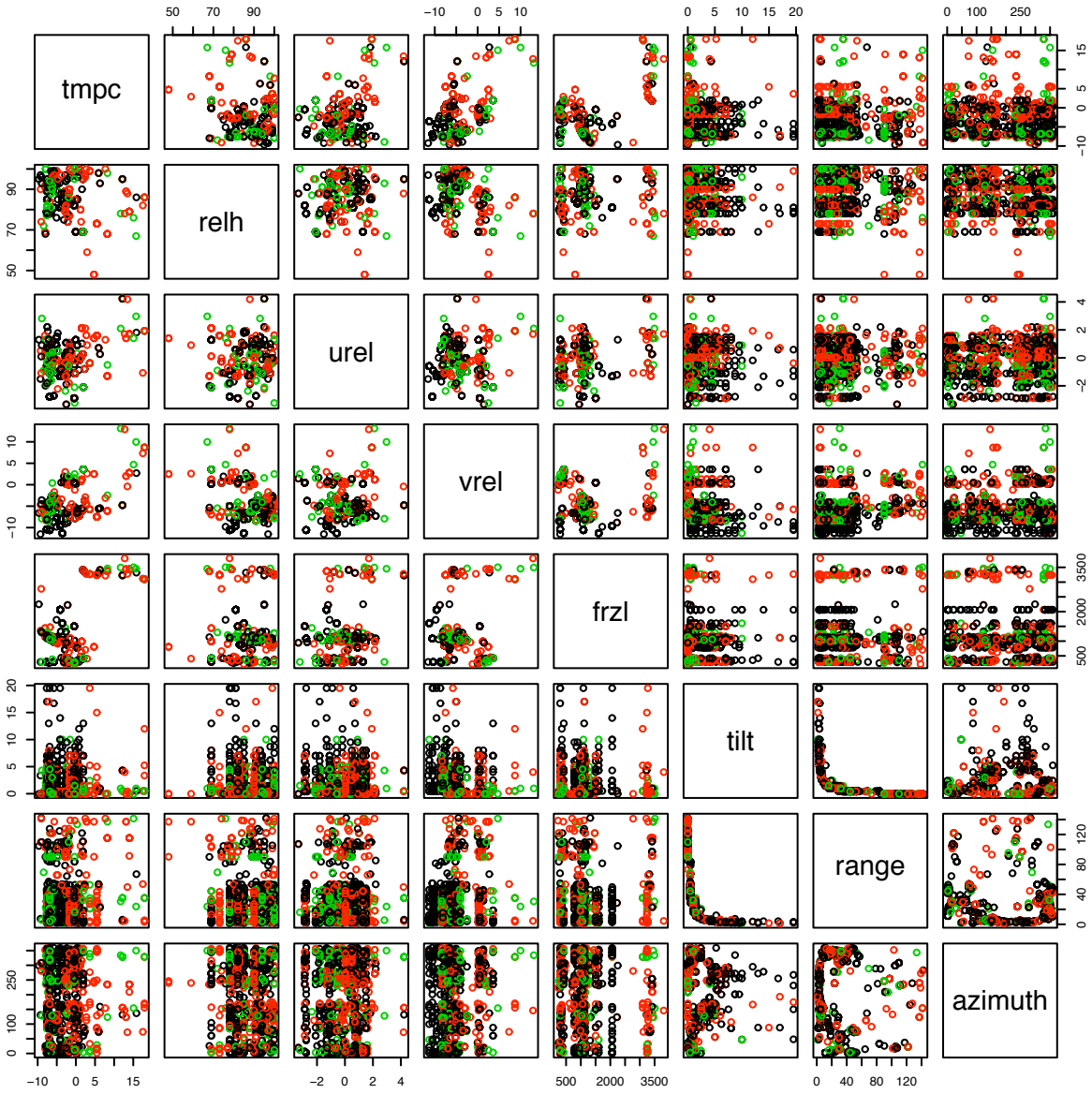

FIG. 2. Pairs plot exploring the relation between variables. While only 6 variables are displayed here, a matrix of all variables was created to study all the relations between variabls. Color codes indicate observed value using the same scale as Figure 1.

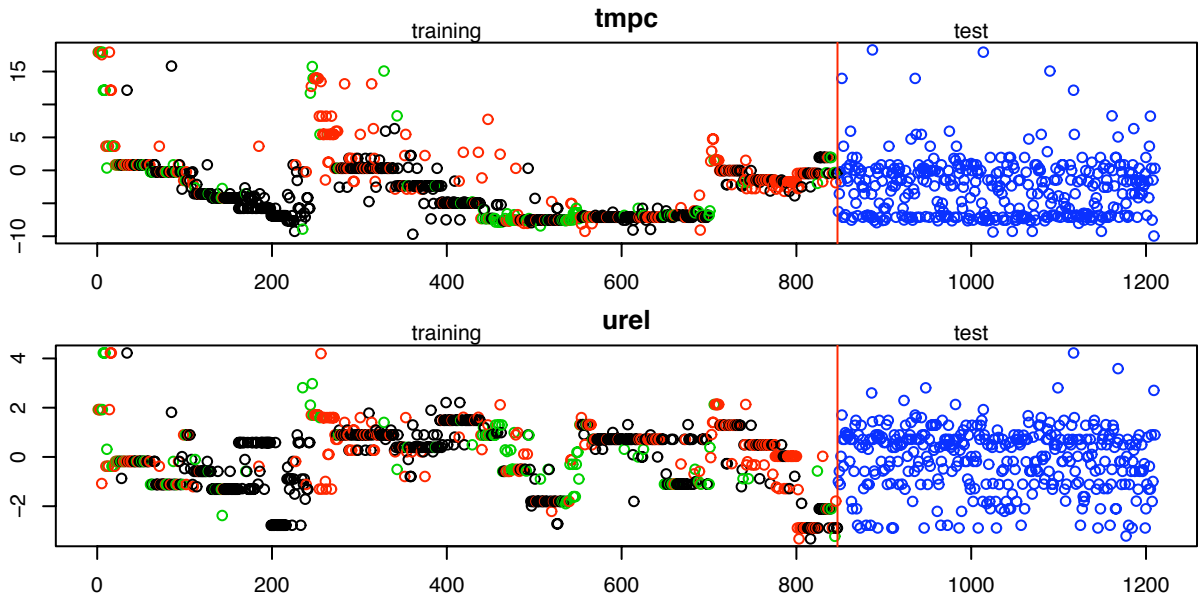

FIG. 3. Pairs plot exploring the relation between variables. Color coding indicates the observed values using same scale as Figure 1.