Artificial Neural Networks for Predicting the Dew Point, Humidity, and the Heat Index

DAVID ADAMS CALLNER[1]
Vice President of Technology Solutions, RIVA Solutions Inc.

ABSTRACT

The National Weather Service (NWS) defines dew point as the temperature to which air must be cooled to reach saturation, assuming air pressure and moisture content are constant. Relative humidity defined by the NWS is the amount of atmospheric moisture present, relative to the amount that would be present if the air were saturated. The heat index is an indicator of what the temperature feels like and more importantly alerts people when dangerous and extremely dangerous conditions exist. These conditions can result in heat cramps or heat exhaustion, and heat stroke. The goal of this study was to use artificial neural networks (ANNs) to predict dew point temperature and humidity to forecast the heat index for the day. This study explores weather data global surface summary of day products produced by the National Centers for Environmental Information (NCEI), to develop general models for dew point temperature and humidity prediction from over 12000 stations that are typically available on any given day dating back to 1929. We explored an aggregative and localized approach with stations for developing our deep learning neural network models. We developed models with input variables that ranged from precipitation, wind gust, mean wind speed, mean visibility, station pressure, sea level pressure, dew point, and mean temperature. We leverage a stochastic approach to the selection of localized stations to run our models against. We were able to achieve very accurate calculations using localized ANNs with two hidden layers of 64 and 32 nodes, respectively. We trained our model with a batch size of 400 with 100 iterations that gave results of the explained variance between 0.91 – 1.0, the mean absolute error of 1.36 – 2.4 degrees Fahrenheit, and the median absolute error of 1.08 – 2.05 degrees Fahrenheit. Predictions of the dew point, humidity and the heat index temperature are particularly important for the agriculture community for preparedness for potential damage of crops, equipment, and death to livestock and humans that affect all communities.

[1] David Adams Callner has a distinguished 20+ year IT career highlights include MITRE, Lockheed Martin, MIT Lincoln Labs, and Raytheon. He has worked as a Technology Advisor for multiple Venture Capital Companies and has launched his own Artificial Intelligence (AI) firms, TuringFantasy, and is the co-founder of OpenCloudCX. Email: dcallner@rivasolutionsinc.com

1. Introduction

Heat stroke results in more than 600 deaths a year in the United States. Far too many children have been inadvertently left in vehicles or have gotten into a vehicle on their own. Vehicular heatstroke tragedies change the lives of parents, families, and communities forever. According to KidsAndCars.org, over 940 children have died in hot cars nationwide since 1990. Even the best of parents or caregivers can unknowingly leave a sleeping baby in a car; and the end result can be injury or even death.

Dew Point and Humidity is most important in the agriculture community because when the temperature is right and there is moisture on the leaves, diseases will grow. For example, diseases will grow on crops and reduce the value of the crops. Vegetable plants may have disease outbreaks on them because there is a pathogen in the area and with occurrence of dew.

Predicting accurate dew point, humidity, and heat index is particularly important for prepardness and awareness of dangerous conditions so parents, families and agriculture communities can take approriate measures to protect our children, livestock, and crops. Researchers have been using mutliple regression analysis since 1990 to compute the heat index.

The computation used for the heat index is a refinement of a result obtained by multiple regression analysis carried out by Lans P. Rothfusz and described in a 1990 National Weather Service (NWS) Technical Attachment (SR 90-23).

For the purpose of this study we will predict the dew point and humidity and use the regression equation of Rothfusz to determine the heat index. The regression equation to tabulated results of Steadman (Journal of Applied Meteorology, 1979) is

$$
\begin{aligned}
HI = & -42.379 + 2.04901523 * T \\
& + 10.14333127 * RH \\
& - .22475541 * T * RH \\
& - .00683783 * T * T \\
& - .05481717 * RH * RH \\
& + .00122874 * T * T * RH \\
& + .00085282 * T * RH * RH \\
& - .00000199 * T * T * RH *
\end{aligned}
$$

Where T is temperature in degrees F and RH is relative humidity in percent.

HI is the heat index expressed as an apparent temperature in degrees F. If the RH is less than 13% and the temperature is between 80 and 112 degrees F, then the following adjustment is subtracted from HI:

$$
ADJUSTMENT = [(13 - RH)/4] * \sqrt{\{[17 - ABS(T - 95.)]/17\}}
$$

Where ABS and SQRT are the absolute value and square root functions, respectively. On the other hand, if the RH is greater than 85% and the temperature is between 80 and 87 degrees F, then the following adjustment is added to HI:

$$
ADJUSTMENT = [(RH - 85)/10] * [(87 - T)/5]
$$

The Rothfusz regression is not appropriate when conditions of temperature and humidity warrant a heat index value below about 80 degrees F.

Dewpoint temperature coupled with relative humidity can be used to determine the amount of moisture in the air. Dewpoint temperature is a good estimate of near-surface humidity, thus the dewpoint temperature can affect the stomatal closure in plants, where low humidity can reduce the productivity of the plants.

Weather predictions use variables like temperature, pressure, wind direction etc. A growing trend of using ANN’s for weather forecasting to predict the dew point, humidity, mean, min, and max tempatures of the atmosphere for a given location, and are made by collecting quantitative data about the current state of the atmosphere and using scientific understanding of atmospheric processes to project how the atmosphere will evolve.

2. Data

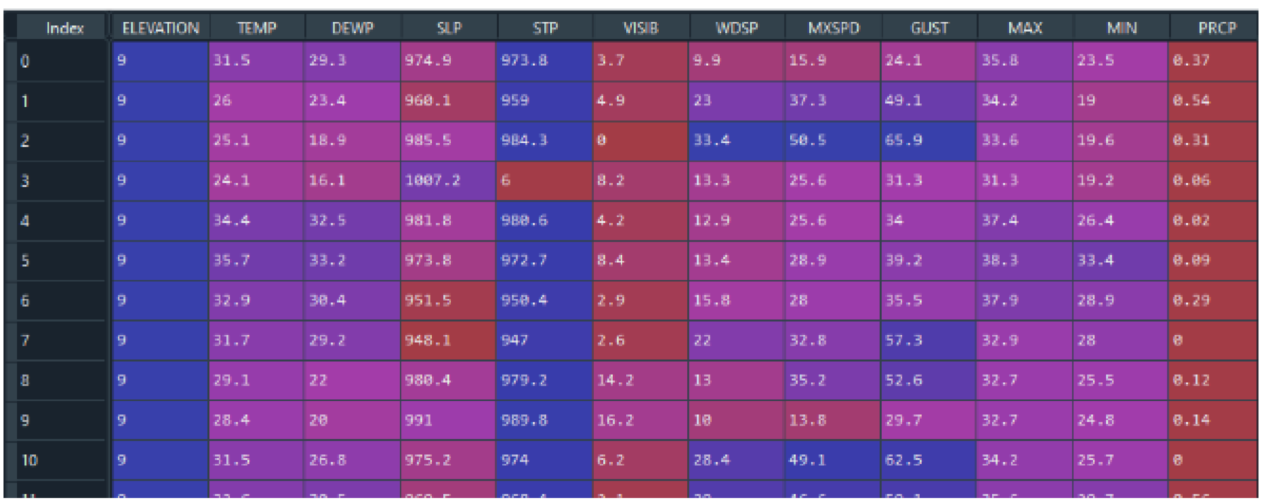
We use NOAA's Global Summary of Day data which is comprised of 18 surface meteorological elements that are derived from the synoptic/hourly observations contained in USAF DATSAV3 Surface data and Federal Climate Complex Integrated Surface Data (ISD). Historical data are generally available for 1929 to the present, with data from 1973 to the present being the most complete.

For some periods, one or more countries' data may not be available due to data restrictions or communications problems. In deriving the summary of day data, a minimum of 4 observations for the day must be present (allows for stations which report 4 synoptic observations/day). Since the data are converted to constant units (e.g, knots), slight rounding error from the originally reported values may occur (e.g, 9.9 instead of 10.0).

The mean daily values are based on the hours of operation for the station. For some stations/countries, the visibility will sometimes 'cluster' around a value (such as 10 miles) due to the practice of not reporting visibilities greater than certain distances. The daily extremes and totals—maximum wind gust, precipitation amount, and snow depth--will only appear if the station reports the data sufficiently to provide a valid value.

Therefore, these three elements will appear less frequently than other values. Also, these elements are derived from the stations' reports during the day, and may comprise a 24-hour period which includes a portion of the previous day. The data are reported and summarized based on Greenwich Mean Time (GMT, 0000Z - 2359Z) since the original synoptic/hourly data are reported and based on GMT.

As for quality control (QC), the input data undergo extensive automated QC to correctly 'decode' as much of the synoptic data as possible, and to eliminate many of the random errors found in the original data. Then, these data are QC'ed further as the summary of day data are derived. However, we expect that a very small % of the errors will remain in the summary of day data.



Index	ELEVATION	TEMP	DEWP	SLP	STP	VISIB	WDSP	MXSPD	GUST	MAX	MIN	PRCP
0	9	31.5	29.3	974.9	973.8	3.7	9.9	15.9	24.1	35.8	23.5	0.37
1	9	26	23.4	960.1	959	4.9	23	37.3	49.1	34.2	19	0.54
2	9	25.1	18.9	985.5	984.3	0	33.4	50.5	65.9	33.6	19.6	0.31
3	9	24.1	16.1	1007.2	6	8.2	13.3	25.6	31.3	31.3	19.2	0.06
4	9	34.4	32.5	981.8	980.6	4.2	12.9	25.6	34	37.4	26.4	0.02
5	9	35.7	33.2	973.8	972.7	8.4	13.4	28.9	39.2	38.3	33.4	0.09
6	9	32.9	30.4	951.5	950.4	2.9	15.8	28	35.5	37.9	28.9	0.29
7	9	31.7	29.2	948.1	947	2.6	22	32.8	57.3	32.9	28	0
8	9	29.1	22	980.4	979.2	14.2	13	35.2	52.6	32.7	25.5	0.12
9	9	28.4	20	991	989.8	16.2	10	13.8	29.7	32.7	24.8	0.14
10	9	31.5	26.8	975.2	974	6.2	28.4	49.1	62.5	34.2	25.7	0

Figure 1. Snapshot of selected variables to train our ANNs

3. Approach

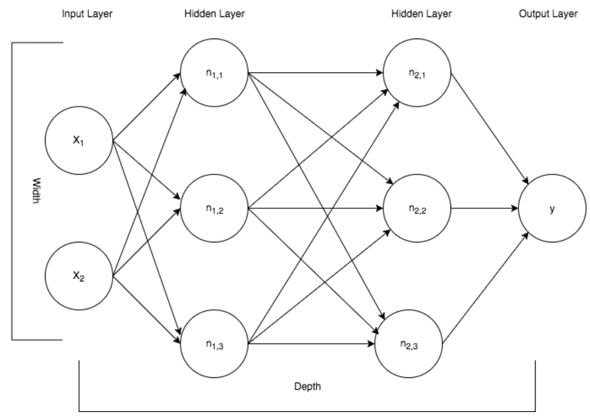
The study started with the use of linear regression models to predict the mean daily temperature for any of the 12135 stations from NOAA’s Global Summary Day data. Linear regression models are extremely powerful and have been used to make numerical, as well as categorical, predictions. However when we attempted to use linear regression to compute the dew point, humidity and the heat index only a few of the input variables (i.e. min, max, and mean tempatures) had a linear correlation. Thus many of the stations explained variance was too low, below .3 in many cases, for this to be a suitable solution.

We then decided to apply neural networks to account for the non-linear variables to determine if we can get suitable explained variance, mean absolute error, and median absolute error to demonstrate that these deep learning models work. Neural networks have a powerful way of utilizing learning techniques based on both linear and non-linear operations.

Neural networks are inspired by biological neurons in the brain which work in a complex network of interactions to transmit, collect, and learn information based off a history of the information that has already been collected. The computational neural networks we are interested in are similar to the neurons of the brain in that they are a collection of neurons (nodes) that receive input signals (numerical quantities), process the input, and transmits the processed signals to other downstream agents in the network. The processing of signals as numerical quantities that pass through the neural network is a very powerful feature that is not limited to linear relationships.

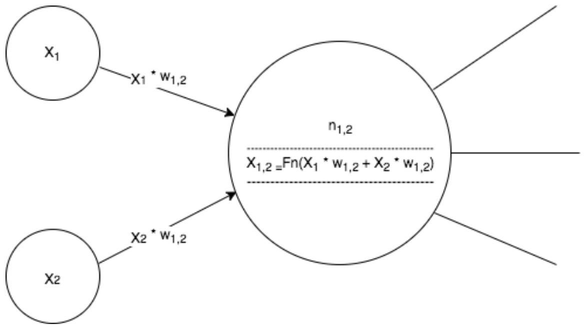
We use supervised learning because the models being trained are built using data that has known dew point, humidity, and heat index target outcomes that the model is trying to learn to predict.

For this approach we use a 4 layer ANN that contains two hidden layers. Graphically, a neural network similar to the one being described in this paper is shown in the image below.



The neural network depicted above contains an input layer on the far left representing two features, x1 and x2, that are feeding the neural network. For the purpose of our study we use all the variables shown in Figure 1 – 12 variables in total. Those two features are fed into the neural network, which are processed and transmitted through two layers of neurons, which are referred to as hidden layers. This depiction shows two hidden layers with each layer containing three neurons (nodes). The signal then exits the neural network and is aggregated at the output layer as a single numerical predicted value.

Each arrow represents a mathematical transformation of a value, beginning at the arrow's base, which is then multiplied by a weight specific to that path. Each node within a layer will be fed a value in this way. Then all the values converging at the node are summed. It is this aggregate of multiplying by weights and summing the products that define the linear operations of a neural network.

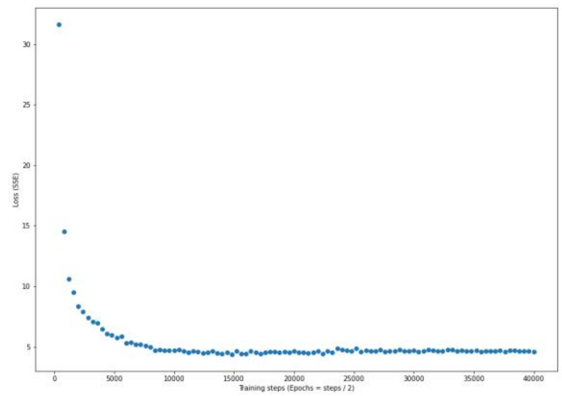


After summation is carried out at each node a special, non-linear, function is applied to the sum, which is depicted in the image above as Fn(...). This special function that introduces non-linear characteristics into a neural network is called an activation function. It is this non-linear characteristic brought about by activation functions that give multi-layer neural networks their power. If it was not for the non-linearity added to the process then all layers would effectively just algebraically combine into one constant operation consisting of multiplying the inputs by some flat coefficient value (ie, a linear model).

Our neural network regressor will iterate over the training data feeding in feature values, calculate the cost function and make adjustments to the weights in a way that minimizes the cost function. We optimize our model by tweaking the hidden layers, number of nodes at each layer, batch size and number of iterations run to build the model.

Model optimization algorithms are very important in building robust neural networks. As examples are fed through the networks architecture (ie, the width and depth) then evaluated against the cost function, the weights are adjusted. The models are learning when the optimizer function identifies that a weight adjustment was made in a way that does not improve (lower) the cost function, which is registered with the optimizer so that it does not adjust the weights in that direction again.

In terms of artificial neural networks, an epoch refers to one cycle through the full training dataset. Usually, training a neural network takes more than a few epochs. In other words, if we feed a neural network the training data for more than one epoch in different patterns, we hope for a better generalization when given a new "unseen" input (test data). Heuristically, one motivation is that it gives the network a chance to see the previous data to readjust the model parameters so that the model is not biased towards the last few data points during training. The graphic below shows our epoch



From the chart above we have not overfitted the model because the evaluation losses never exhibit a significant change in direction toward an increasing value. With these model tweaks we then stocastically selected a few stations to run our model againt.

4. Results

Explained variance (also called explained variation) is used to measure the discrepancy between a model and actual data. In other words, it is the part of the model's total variance that is explained by factors that are actually present and isn't due to error variance.

Higher percentages of explained variance indicates a stronger strength of association. It also means that you make better predictions (Rosenthal & Rosenthal, 2011).

$$r2 = R2 = \eta 2$$

Explained variance can be denoted with r2. It is called eta squared (η2) and in regression analysis, it's called the Coefficient of Determination (R2). The three terms are basically synonymous, except that R2 assumes that changes in the dependent variable are due to a linear relationship with the independent variable; Eta2 does not have this underlying assumption.

Explained variance is calculated with the "eta-squared (η2)" ratio Sum of Squares (SS)between to SStotal; It is the proportion of variances for between group differences.

R2 in regression has a similar interpretation: what proportion of variance in Y can be explained by X (Warner, 2013).

The Mean Absolute Error (MAE) is the average of all absolute errors. The formula is:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n} |x_i - x|$$

Where:

n = the number of errors,
Σ = summation symbol (which means "add them all up"),
|xi – x| = the absolute errors.

The formula may look a little daunting, but the steps are easy:

- Find all of your absolute errors, xi – x.
- Add them all up.
- Divide by the number of errors. For example, if you had 10 measurements, divide by 10.

The median absolute error is particularly interesting because it is robust to outliers. The loss is calculated by taking the median of all absolute differences between the target and the prediction. If ŷ is the predicted value of the ith sample and yi is the corresponding true value, then the median absolute error estimated over n samples is defined as follows:

$$MedAE(y, \hat{y}) = median(|y_1 - \hat{y}_1|, \ldots, |y_n - \hat{y}_n|)$$

In the table below we display our results of our model against 5 stochastically chosen stations to calculate the dew point, humidity, and the heat index.

Station / Measurement	ORLANDO INTERNATIONAL AIRPORT, FL	LEE CO AIRPORT, VA	MANITOWISH WATERS AIRPORT, WI	CORAL HARBOUR, CA	CAMBRIDGE BAY, CA
The Explained Variance	0.91	.93	.96	.99	1.0
The Mean Absolute Error	2.40 degrees Fahrenheit	2.05 degrees Fahrenheit	1.66 degrees Fahrenheit	1.83 degrees Fahrenheit	1.57 degrees Fahrenheit
The Median Absolute Error	2.05 degrees Fahrenheit	1.93 degrees Fahrenheit	1.88 degrees Fahrenheit	1.34 degrees Fahrenheit	1.08 degrees Fahrenheit

5. Conclusion

Dew point temperature and humidity proved to be good for ANN estimation with the selected stations that we used. Accurate heat index was calculated following the inputs of the estimated humidity. We discovered vast differences in the models with different hidden layers, batch size, and iterations. We found the highest explained variance with 64, 32 hidden nodes in a 4-layer ANN with 400 batch size and 100 iterations. The efficiency of the ANN is dependent on the optimal tuning for the input parameters and training algorithm employed. Each station calculation took about 20 minutes to run fully which limited the number of stations that we could test against. The evaluation of the final models with weather data from 5 separate stations and for 2020 year showed that the dew point and humidity predictions had mean absolute errors (MAEs) of 2.40°, 2.05°, 1.66°, 1.84° and 1.57°F, respectively. These predictions are useful for decisions in agriculture because dew point, humidity and the heat index temperature are particularly important for the agriculture community for preparedness for potential damage of crops, equipment, and death to livestock and humans that affect all communities.

6. Future Work

To expand on this research, it is imperative to test against more stations and past years. We will be working with vendors who provide supercomputing capabilities to scale this work.

Acknowledgments. This work was funded by RIVA Solutions, Inc. through the Innovation Lab. To foster innovation, RIVA invests in building new capabilities in key technical research and development, bringing the successful outcomes together in an array of integrated emerging technology solutions platforms through our Innovation Lab. The Innovation Lab consists of a group of external thought leaders that continually challenges the markets thinking to solve the Government's toughest problems and potential solutions.

REFERENCES

Hubbard, K. G., R. Mahmood, and C. Carlson, 2003: Estimating daily dew point temperature for the northern Great Plains using maximum and minimum temperature. Agron. J., 95, 323–328.

Kimball, J. S., S. W. Running, and R. Nemani, 1997: An improved method for estimating surface humidity from daily minimum temperature. Agric. For. Meteor., 85, 87–98.

Maqsood, I., M. R. Khan, and A. Abraham, 2004: An ensemble of neural networks for weather forecasting. Neural Comput. Appl., 13, 112–122.

Mittal, G. S., and J. Zhang, 2003: Artificial neural network-based psychrometric predictor. Biosyst. Eng., 85, 283–289.