Predict Disease Spread Development Guide

Oscar Hernandez

Northwestern University

June 9th, 2019

Predict Disease Spread Development Guide

This model development guide is focused on covering the steps taken to train several models using time series data. The data are provided by several U.S. Federal Government agencies such as the Centers for Disease Control and Prevention. The primary goal is to make use of historical time series data along with features that provide various descriptive measures of the observations to achieve accurate forecasts.

Specifically, the data consist of weekly total cases of Dengue Fever for San Juan, Puerto Rice and Iquitos, Peru. The date ranges that were used to train the models vary for each city. San Juan's dates are from April 1990 to April 2008. Iquitos dates are from July 2000 to June 2013. The weekly total cases of Dengue Fever forecasts for San Juan and Iquitos will be for the next five and three years, respectively. Furthermore, the data include features such as the maximum temperature, total precipitation and vegetation information.

One of the challenging parts of building the time series models was gaining background knowledge on the topic and understanding how to effectively utilize the data available. Without much domain knowledge on disease spread, it wasn't apparent which features were going to be strong predictors of the total cases of Dengue Fever. Furthermore, selecting models that can incorporate predictor variables along with the time series data was another important part of creating the development guide. Overall, completing multivariate time series analysis and modeling is what made this work challenging.

This exercise is significant because of the potential implications that result from disease spreading across the world. Disease spread has significant economic and health costs to society which, if not recognized in a timely manner, can lead to irreversible outcomes. Therefore, governments and various health agencies should be making use of the analytical tools available to spot trends and make accurate forecasts. Arguably, any health organization that is not making use of forecasting from the data available to them is being negligent. Although there exists plenty of examples that utilize univariate time series data effectively, capturing the complexity of disease spread is best served using multivariate time series data. Much of the data that is available today includes many features that can be useful in increasing the predictive accuracy of a model. Therefore, it makes sense to at least consider using the multiple features that are available if it results in better model performance. Overall, the work involved in this development guide shows that it is possible to make effective multivariate time series models using environmental data.

Literature Review

A significant amount of academic and industry research has been done to better understand the relationship that environmental factors have on disease spread. Typical environmental variables include rainfall and air temperature. Furthermore, the advancement of satellite imagery has allowed researchers to incorporate new variables such as crop growth across various regions. Many of those studies utilize time series data along with various statistical and machine learning model classes to generate forecasts of disease spreads.

Imai, Armstrong, Chalabi, Mangtani and Hashizume (2015) studied the association of air pollution and weather with mortality using time series regression models. Specifically, these researchers utilized a Poisson model and incorporated the change in immunity into their model development. Bhaskaran, Gasparrini, Hajat, Smeeth and Armstrong (2013) wrote a peer reviewed article that outlined the advantages and disadvantages of using a Poisson regression model to conduct environmental epidemiology research. The article was not necessarily a study

of a specific disease or health condition but rather a framework that researchers can follow when attempting to use time series data in their research.

Morral-Puigmal, Martínez-Solanas, Villanueva and Basagaña (2018) used 17 years' worth of daily weather data and gastroenteritis hospital admissions in Spain to study the effects of various weather variables. They concluded that hot and cold temperatures were associated with an increased risk in gastroenteritis and that precipitation had a protective impact on the illness. The researchers used daily time series data across various regions to develop quasi-Poisson regression models. Lal, Ikeda, French, Baker and Simon Hales (2013) utilized Seasonal Auto Regressive Integrated Moving Average (SARIMA) models to investigate the association between monthly climate change and enteric diseases (campylobacteriosis, salmonellosis, cryptosporidiosis and giardiasis) in New Zealand. Their results suggested that "temporally lagged relationships between climate variables and national communicable disease incidence data can contribute to disease prediction models and early warning systems" (Lal et al., 2013).

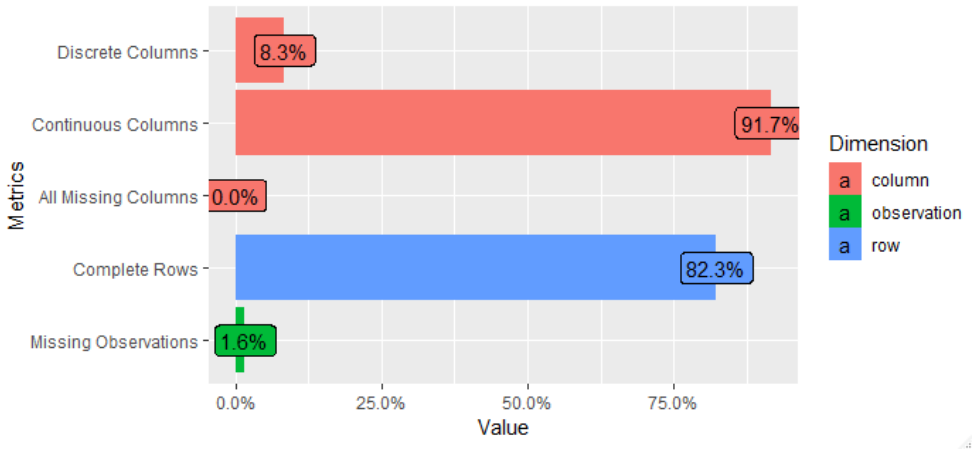
Song et al. (2015) built a SARIMA model by incorporating eight climate variables along with hand foot mouth disease (HFMD) time series data. Their research concluded that HFMD and climate patterns are strongly correlated.

These studies provided insight into previous research involving multivariate time series modeling with respect to understanding the relationship with disease spread and environmental factors. It is reassuring to know the diversity of modeling classes that are applicable to this topic.

Data Preparation

The data that was obtained for this model development guide was relatively clean in the sense that there weren't many issues that needed to be fixed. Figure 1 displays the dimensions from the training data set. The first takeaway is that most of the variables are continuous which is suitable for the purpose of developing regression models. Furthermore, there appears to be minimal data that is missing. Although not presented in this guide, the same plot was created for the test data set. The results were similar to the training set. Overall, this initial review of the data sets indicates that there will not be much data preparation needed.

Figure 1: Training Data Summary



Although certain model classes can handle missing values during the training process, it was decided to still impute the missing values in both data sets. The imputations were done using the R "na.locf" function which allows for the last value before a missing value to be chosen as the replacement for the missing value. Furthermore, both the training and test data sets were

combined prior to making this change in order to get the most accurate results. Table 1 displays the summary statistics for the entire data set (except the categorical variables) after the missing values were fixed. There doesn't appear to be any erroneous values based on the ranges presented in the table. For example, if the *station_avg_temp_c* variable would have included a value of 88, it would be a reason for concern since it most likely is an error given that the highest temperature ever recorded on Earth is 56.7 Celsius. These types of errors (e.g. errors or outliers) have a negative impact on regression models.

Since the goal is to forecast the total cases of Dengue Fever in both San Juan and Iquitos, the decision was made to split the data based on the *city* variable. Therefore, two models – one for each city – will be trained for each model class that is chosen. After obtaining the Juan and Iquitos data sets, a time series object was created for both sets using the R "ts" function. The time series objects were developed with the following parameters:

- Data = *total_cases* variable (e.g. san_juan.train$total_cases)
- Frequency = 52 (number of weeks in a year)
- Start Date (San Juan) = 4/30/1990; Start Date (Iquitos) = 7/2/2000

Table 1: Summary Statistics for Default Variables (Combined Data Set)

Statistic	N	Mean	St. Dev.	Min	P(25)	Median	P(75)	Max
year	1,872	2,003.19	6.29	1,990	1,999	2,004	2,008	2,013
weekofyear	1,872	26.49	15.01	1	13.8	26	39	53
ndvi_ne	1,872	0.13	0.15	-0.46	0.03	0.11	0.24	0.51
ndvi_nw	1,872	0.13	0.13	-0.46	0.04	0.11	0.22	0.65
ndvi_se	1,872	0.20	0.07	-0.02	0.15	0.20	0.25	0.54
ndvi_sw	1,872	0.20	0.09	-0.06	0.14	0.19	0.25	0.55
precipitation_amt_mm	1,872	44.04	42.16	0.00	9.12	36.50	67.36	390.60
reanalysis_air_temp_k	1,872	298.72	1.38	294.55	297.68	298.61	299.89	302.20
reanalysis_avg_temp_k	1,872	299.25	1.27	294.89	298.26	299.29	300.27	303.33
reanalysis_dew_point_temp_k	1,872	295.28	1.53	289.64	294.16	295.66	296.50	298.45
reanalysis_max_air_temp_k	1,872	303.46	3.21	297.80	301.00	302.50	305.50	314.10
reanalysis_min_air_temp_k	1,872	295.72	2.60	286.20	293.80	296.20	298.00	299.90
reanalysis_precip_amt_kg_per_m2	1,872	40.56	44.61	0	12.5	27.0	53.1	570
reanalysis_relative_humidity_percent	1,872	82.24	7.21	57.79	77.21	80.31	86.80	98.61
reanalysis_sat_precip_amt_mm	1,872	44.04	42.16	0.00	9.12	36.50	67.36	390.60
reanalysis_specific_humidity_g_per_kg	1,872	16.78	1.55	11.72	15.59	17.13	18.03	20.46
reanalysis_tdtr_k	1,872	4.95	3.54	1.36	2.36	2.87	7.76	16.03
station_avg_temp_c	1,872	27.22	1.27	21.40	26.36	27.41	28.18	30.80
station_diur_temp_rng_c	1,872	8.06	2.22	4.04	6.37	7.24	9.75	15.80
station_max_temp_c	1,872	32.47	1.95	26.70	31.10	32.80	33.90	42.20
station_min_temp_c	1,872	22.14	1.62	14	21.1	22.2	23.3	27
station_precip_mm	1,872	38.28	44.93	0	8.9	23.8	52	543
total_cases	1,872	0.00	0.00	0	0	0	0	0

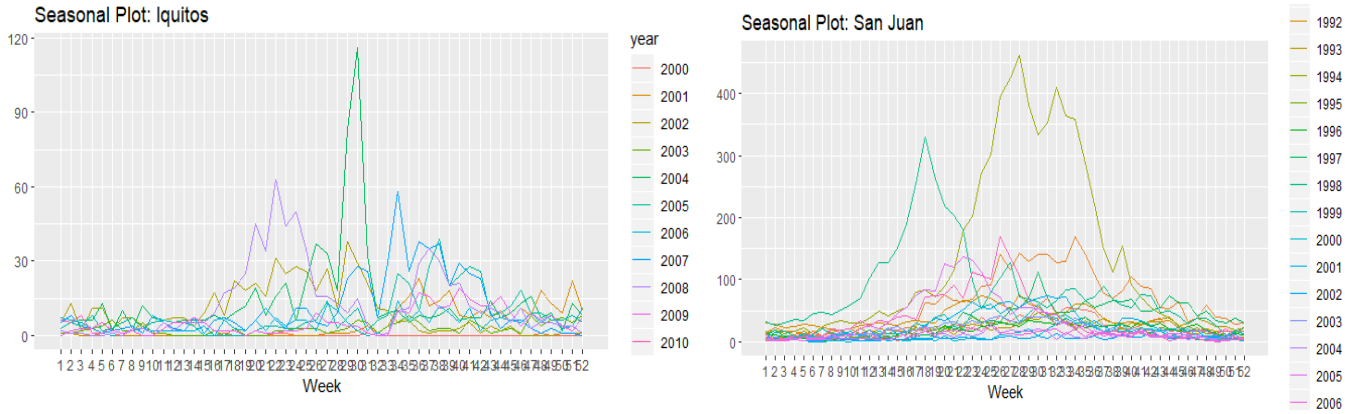
Exploratory Data Analysis

Several steps were taken to carefully explore the time series objects created. First, an Augmented Dickey-Fuller Test was conducted on both objects to test for stationarity using an

alpha value of 0.05. Both time series objects had a p-value of 0.01 which means the results were statistically significant. This suggests that the time series data do not have a time dependent structure. This is an important takeaway because it provides evidence to suggest that incorporating the environmental variables would be useful (i.e. more accurate forecasts).

Based on the key learnings from the Literature Review, it is reasonable to assume the time series data associated with San Juan and Iquitos will have seasonality. Figure 2 displays the seasonal plots for both time series objects.

Figure 2: Season Plots



The Iquitos plot indicates that Dengue Fever cases occur during the Winter season (May through September) which is Peru's driest season. This information is useful since health organizations would want to further study why this seasonal pattern exists and if it is changing. A change in the seasonal pattern might indicate that the disease spread might be growing (i.e. summer months include more cases). The San Juan plot shows a noticeable spike in 1993. However, the number of total cases appears to be decreasing based on the 2004 – 2006 data points. This information is important because it means that any model fitted on the data most likely wouldn't forecast a higher number of cases in subsequent years. Overall, there does appear to be a seasonal pattern of Dengue Fever cases for both cities.

Model Development

Three model classes were chosen to forecast the *total_cases* dependent variable. The test data provided requires a 5- and 3-year weekly forecast for San Juan and Iquitos, respectively. Given the seasonality in the data, a Seasonal and Trend decomposition using Loess (STL) model was chosen which is useful for estimating nonlinear relationships. The R "mstl" function was utilized to create the model. Based on the Literature Review, a seasonal ARIMA model with multiple regressor variables was developed using the R "auto.arima" function. The last model that was developed using the R "nnetar" function was a neural network with multiple regressor variables. The neural network model class allows for complex nonlinear relationships between the regressor variables and the dependent variable, *total_cases*. The multiple regressors variables that will be in the ARIMA and neural network are as follows:

- *station_max_temp_c* – Maximum temperature
- *station_min_temp_c* – Minimum temperature
- *station_avg_temp_c* – Average temperature

- *station_precip_mm* – Total precipitationstation_diur_temp_rng_c – Diurnal temperature range

Although there are several other variables in the data set, the decision was made to utilize only a subset as a test to determine if accurate forecasts can still be attained. Overall, the expectation is that these environmental variables will increase the predictive ability of the models using them. In other words, the STL model should have the less accurate forecasts.

Model Comparison

The final step is to compare each of the model's predictive accuracy. Table 2 displays the accuracy metrics used to measure performance on the training sets for San Juan and Iquitos. The model with the lowest MAE will be considered the best performing one when considering predictive accuracy. The Neural Network model had the highest predictive accuracy in every accuracy metric for both cities. Its RMSE and MAE values were almost 50% better than the other two models' metrics. The Neural Network model seems to have been able to capture the nonlinear relationship with the environmental variables and *total_cases*. Furthermore, the ARIMA model was the worst performing model which is surprising since it incorporated the environmental variables and the STL model did not. Given the significant difference in performance results, it's clear that the Neural Network model is the best one which will then be used to make forecasts on the test set.

Table 2: Final Models Performance Results

	San Juan			Iquitos		
Model	**RMSE**	**MAE**	**MASE**	**RMSE**	**MAE**	**MASE**
STL	12.20	7.63	0.21	6.48	3.83	0.41
Neural Network	6.48	4.81	0.13	2.85	2.15	0.23
ARIMA	13.36	8.04	0.22	6.91	3.90	0.41

Conclusion

Limitations

There were several limitations that impeded the performance of the models. The first issue is that there wasn't any feature engineering completed to create new variables that may have contributed to higher predictive accuracy. Also, not all the variables were used to train the Neural Network and ARIMA model. The data set included a variable pertaining to vegetation changes for each city which could have also improved the forecasts. Overall, the limitations were around variable selection. However, it's important to note that simply using extra variables can lead to collinearity issues and/or overfitting the data.

Future Work and Key Learnings

The analysis completed provides further statistical affirmation that a strong relationship between health conditions and environmental factors exists. From a forecasting perspective, the existence of this relationship is important since health organizations can take comfort in knowing that any future research is worth pursuing. Another key takeaway is that using a model class that can capture complex nonlinear relationships with a dependent variable such as a neural network is beneficial. Any future work to improve the models presented in this guide should involve engineering new features that represent environment changes (e.g. average change in precipitation) and consider using more advanced modeling approaches (e.g. gradient boosting).

References

Imai, C., Armstrong, B., Chalabi, Z., Mangtani, P., & Hashizume, M. (2015). Time series regression model for infectious disease and weather. *Environmental Research, 142,* 319-327.

Bhaskaran, K., Gasparrini, A., Hajat, S., Smeeth, L., & Armstrong, B. (2013). Time series regression studies in environmental epidemiology. *International Journal of Epidemiology, 42*(4), 1187-1195.

Morral-Puigmal, C., Martínez-Solanas, È., Villanueva, C.M., & Basagaña, X. (2018). Weather and gastrointestinal disease in Spain: A retrospective time series regression study. *Environment International, 121*, 649-657.

Lal, A., Ikeda, T., French, N., Baker, M.G., & Simon Hales. (2013). Climate variability, weather and enteric disease incidence in New Zealand: Time series analysis. *PLoS ONE, 8*(12), E83484.

Song, Y., Wang, F., Wang, B., Tao, S., Zhang, H., Liu, S., Ramirez, O., & Zeng, Q. (2015). Time series analyses of hand, foot and mouth disease integrating weather variables. *PLoS ONE, 10*(3), E0117296.