Abstract

Dengue fever is a mosquito-borne infection that affects almost 250 million people every year. Being able to predict where outbreaks occur beforehand can save tens ouf thousands of lives per year. Past research has shown that mosquitos thrive in humidity. This paper fits an inferential, multiple linear regression model to estimate that a one-unit increase in specific humidity will increase dengue fever cases by 146% (30%, 380%). World health officials should use this information to expand efforts to limit mosquito populations in humid areas.

Keywords

Dengue fever, multiple linear regression, humidity, inference

Contents

1 Introduction

As one of the most prolific diseases in the world, dengue fever is a mosquito-borne infection that affects almost 250 million people every year (Guzman 2010). Mainly endangering the tropics, the fever at best causes symptoms like vomiting and at worst can turn life-threatening. Thousands die every year while researchers attempt to find the best way to fight the pandemic. There currently is no fully-effective vaccine.

The disease first started to become prevalent after World War II as mosquitoes were able to travel around the world much easier than before (WHO 2015). Since the disease can't spread directly between people, the spread of mosquitoes led to the spread of dengue fever. Specifically, the Aegypti mosquito is the main vector for the virus. Researchers understand this link and have turned their attention to efforts like eliminating still water where mosquitoes breed and encouraging residents to wear clothes that cover more of their skin. If you want to stop dengue fever, you have to stop the mosquitoes. This link has been known for decades (Henchal 1990).

The burder of dengue fever is likely to continue to increase over the coming decades as the pace of urbanization increases and the effects of climate change become more present (Fatima 2018). Dwindling water supplies, environmental change, and poor sanitation will be factors that make the need for proper dengue fever management more critical.

Managing a viral disease like dengue fever requires careful coordination. Many underlying factors lead to small outbreaks which can exponentially increase in size unless stopped. Since dengue fever is common in over 110 countries, finding where the next hotspot of incidences can be challenging. Cities can have huge breakouts for a few weeks, but then immediately go back to only a few cases.

Global health officials such as the WHO have processes in place once they arrive at a location to quell an outbreak, but in recent years they have become better at predicting which location they will need to be at before an outbreak happens. Resources are limited and local facilities can't keep excess stock of vaccines because of the costs. Being predictive rather than reactive can mean response times that are days and weeks better (WHO 2015). This small improvement can be the reason thousands more live every year and millions don't become infected at all. The worst versions of dengue fever kill fairly quickly so timeliness is critical.

The place of big data in this problem is to inform health officials of which factors can be used to predict likely hotspots of dengue fever before they happen. The viral process involves the movements of people, mosquitos, and resources, all processes that mathematical models are becoming more equipped to handle over the years. Researchers have tried hard to model where incidences will occur so they can send vaccines early and treat people

promptly. A proper model can save lives by directing resources where they need to be.

As mentioned before, mosquitoes are the culprits of dengue fever. Dengue fever goes where mosquitoes thrive and mosquitoes thrive in humid areas. Research has shown that humidity, particularly specific humidity, is the best predictor of mosquito population size (Sintorini 2018). Given both of these things, it makes sense to say that specific humidity levels can predict dengue fever cases in a city.

To test this, this paper seeks to answer this research question: all else being equal, what is the effect of a one unit change in humidity levels on predicted cases of dengue fever?

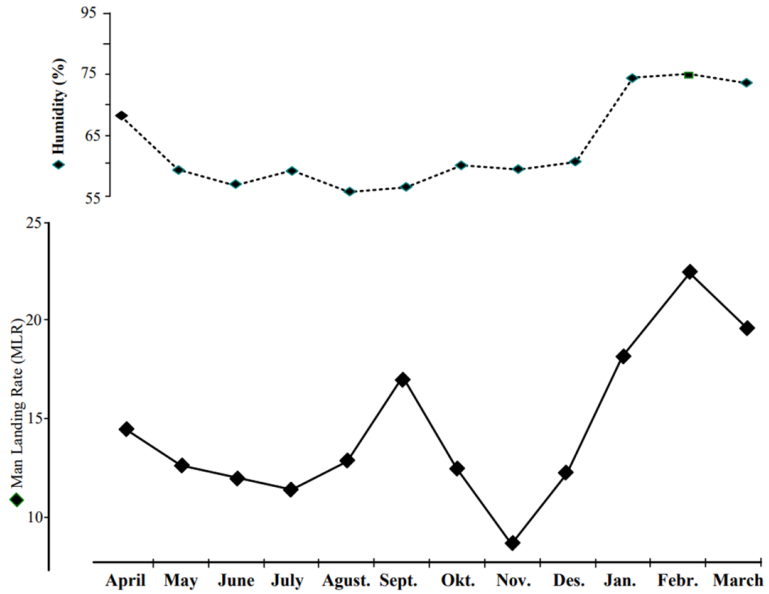
2 Related Work

2.1 Sintorini 2018: Humidity and Mosquitos

Due to recent weather changes in South East Asia, dengue fever has been on the rise in 2018. Sintorini, an ecologist, conducted an observational study in Jakarta, Indonesia to gather data on mosquito populations and then relate that data to environmental data such as humidity. The objective of the study was to help find early warning signs of the spread of dengue fever.

In his study, the measure of mosquito population size was Man Landing Rate (MLR). MLR is a common ecological measure and can simply be explained as a proxy for mosquito population size. The below figure shows the relationship between humidity and MLR for each money of the study.

Figure 1



These two time series have a significant correlation with a p-value of less than 0.01%. The methodology of the paper was only the correlation coefficient. This paper will attempt to expand on this work using a more involved linear model that includes confounders.

2.2 McMurren et. al. 2013: Climate and Dengue Fever

Since 2009, Paraguay has been struggling with dengue fever more so than its neighboring countries. With a lack of a coordinated effort in the country, the Paraguay government opened up many forms of data for researchers to use to help detect dengue fever outbreaks.

McMurren et. al. worked in 2013 to build a data-driven model that incorporated climate, cartography, and population data. Their work was meant to be a proof of concept of how the opening of data by the government could be useful in prediction. The reserachers built an online tool that could be used over the coming years as long as the needed input data remained available.

With an accuracy of 94.78%, their binary prediction model was seen as a success. While the binary definition of an "outbreak" was in question, this serves as further evidence that climate data including humidity can be used to predict dengue fever cases.

2.3 Boussion 2012: Climate and Dengue Fever

During a similar time as the previous research paper, Bouisson was working to use climate data specifically to predict dengue fever cases. Researchers analysed epidemiological and climatological data in Noumea, New Caledonia to find correlations. This allowed them to use further statistical models to draw inferences. Noumea has reliably kept data on dengue fever which enabled this research to occur.

In their first inferential model, researchers used temperature and humidity only to accurately predict outbreaks. They found that hot and dry days in the winter led to outbreaks while cool and wet days in the summer led to outbreaks. This reveals a "sweet-spot" for mosquitos to thrive in.

In their second model, prediction was the focus so more variables were included to improve accuracy at the sake of interpretability. This model is currently used by public health authorities to manage the disease.

In summary, all past research points to humidity and dengue fever being intimately related. This paper will attempt to replicate these findings in a separate dataset.

3 Conceptual Framework

3.1 Multiple Linear Regression (MLR)

A multiple linear regression (MLR) model is a statistical model that relates multiple predictor variables to a continuous response variable. The model attempts to explain the relationship between the variables in a linear fashion. The stable method of least squares is used to estimate coefficients and fit a hyperplane through the data.

The MLR equation describes how the expected response variable changes due to one-unit changes to predictor variables. Different from single-variable linear regression, the coefficeints in a MLR model are interpreted as partial slopes. This means that a beta coefficient in a MLR model is itnerperted as the effect of a one-unit increase in the predictor variables, *holding all other variables constant*. This is the conditional part that makes it a partial slope rather than an overall slope. MLR models take account of confounding relationships and are simplistic enough for inference.

3.2 Inference vs. Prediction

Models can be built for two main reasons, inference and prediction. Inferential models attempt to tease out cause and effect. Great care is put into what variables are put into a model. All possible confounding relationships must be included to obtain a causal interpretation of a partial slope in a MLR model. A confounding variable is a third variable that is related both to the response variable and the independent variable in question. Parsimony wins out in infernential models as unnecessarily included variables increase variance and add noise. Simplistic, linear models are best for inference

Predicitve models need much less care and parsimony. They can be much more complex and be 'black boxes' as long as out-of-sample predictive accuracy increases. The research question of this paper is inferential so the former style of model will be built. Each variable will be investigated in depth before being added.

3.3 Climate Data

While there are many different climate measures in the full dataset, this section will explain the measures that are eventually used in the model. Explanations of all variables in the dataset are included in Appendix A.

Specific humidity is an absolute measure of humidity. It is defined as grams of water per kilogram of air. In this paper, specific humidity will be the average recorded value for a week. Relative humidity is another measure that takes into account the current temperature to calculate how much of

the possible maximum amount of water is in the air. Since this measure depends upon another variable, it is not used in this paper.

Dew point is the temperature to which air must be cooled to become saturated with water vapor. When the temperature reaches this point, water can condense on objects like grass creating a "morning dew." The average weekly value recorded in Kelvin is what is used in this dataset.

4 Methodology

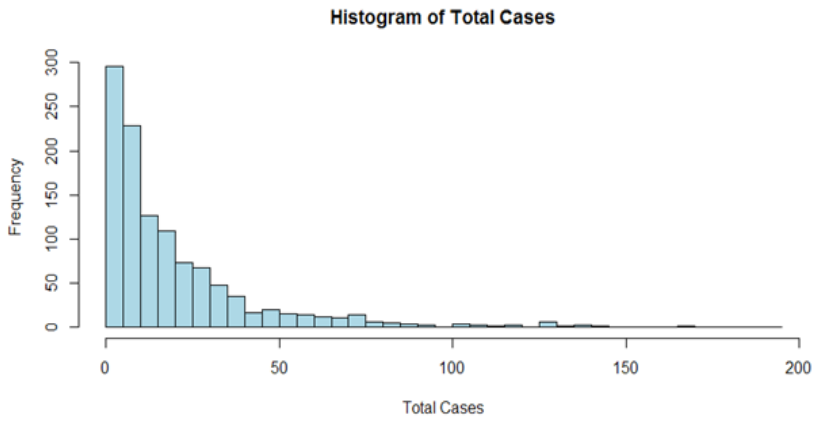
4.1 Data Source

This analysis uses a dataset sourced from an online competition hosted by DrivenData. The dataset includes weekly information from two Latin American cities, San Juan, Puerto Rico and Iquitos, Peru. Along with the number of dengue fever cases each city encountered in the week, included are various environmental measures that describe precipitation, temperature, vegetation, and more. The variables are all intended to be in some way related to how mosquitoes breed and spread. The dataset includes complete data on over 1100 weeks from these two cities in years between 1990 and 2010.

The climate data in the dataset is not the reported values, but actually a reanalysis of the initial readings. A very common practice, reanalysis is where climate scientists can look back on a wide range of time and compute values under a single model. Over the twenty year span of the dataset, climate stations in the cities invariably changed models or methods for calculating values like average temperature so reanalysis goes back and standardizes the model used. Variation from modeling choices is removed and best practices are instilled. A listing of all variables is included in Appendix A.

4.2 Model Building

Figure 2



The histogram in Figure 2 shows the distribution of weekly dengue fever cases for all points in the dataset. The distribution is skewed right with approximately 90% of weeks having less than 50 cases. Some highly active weeks are omitted from the histogram, but reach up to 329 cases. Areas have huge breakouts, but mostly have weeks with low incidences. Because of the skew in total cases, log(total.cases) will be the response variable for the model. As well, mosquito-borne fevers have a natural inclination towards an exponential distribution so a log transformation makes theoretical sense. Cases spread faster the more people in an area have the fever.

Figure 3

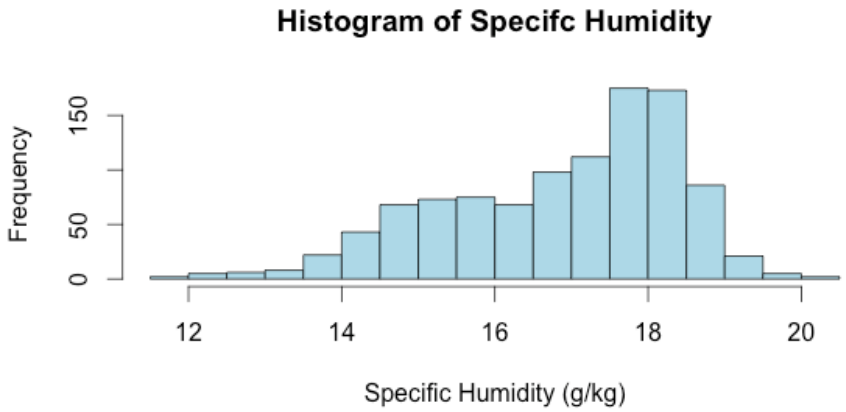


Figure 3 above is a histogram of recorded specific humidity levels in the dataset. The distribution has all values between 11 and 21 grams of water per kilogram of air and a median at 17 g/kg. As well, the distribution is left skewed. Overall, these are very humid weeks since the dataset is sourced from Peru and Puerto Rico.

To find a proper estimate of the partial slope between specific humidity and log(total.cases), all confounders need to be included in the model. Figure 4 below shows comparative boxplots of the relationship between city and log(total.cases) on the left and city and specific humidity on the right. On the left, the median of San Juan is significantly higher than Iquitos. San Juan has on average more cases of dengue fever. On the right, San Juan has a median significantly lower than Iquitos. Iquitos has more humid weeks on average than San Juan. Because the city variable is correlated with both our response and predictor variable, it is a confounder and will be included in our model.

Figure 4

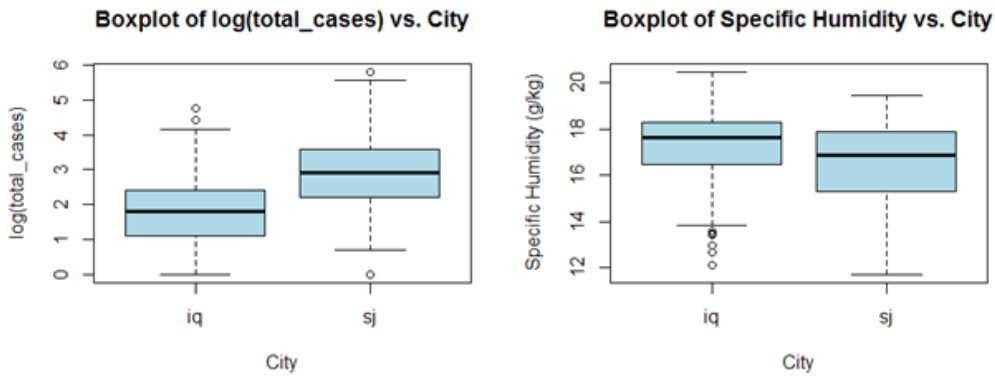


Figure 5 again shows comparative boxplots, but now of relationship between season and log(total.cases) on the left and season and specific humidity on the right. On the left, Fall has on average the most cases of dengue fever while spring has significantly less. Summer and Winter differ from both Fall and Spring with a median in the middle. On the right, Fall and Summer have more humid weeks than Spring and Winter. Because the season variable is correlated both with specific humidity and log(total.cases), it is a confounder and will be included in the model.

Figure 5

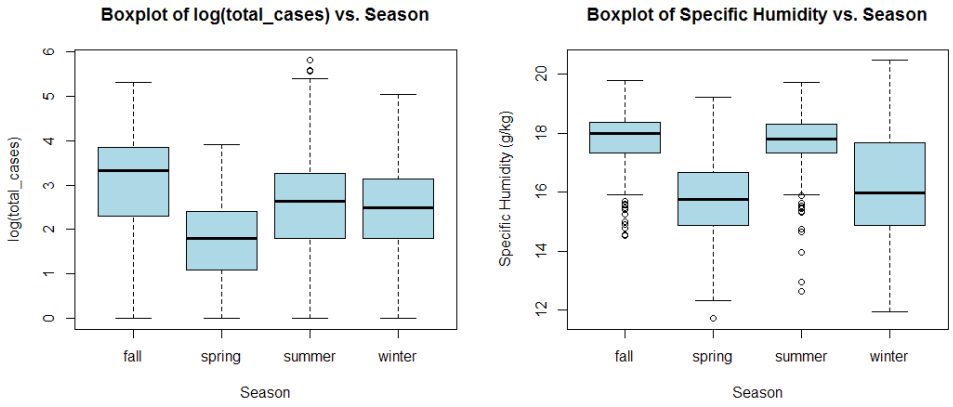
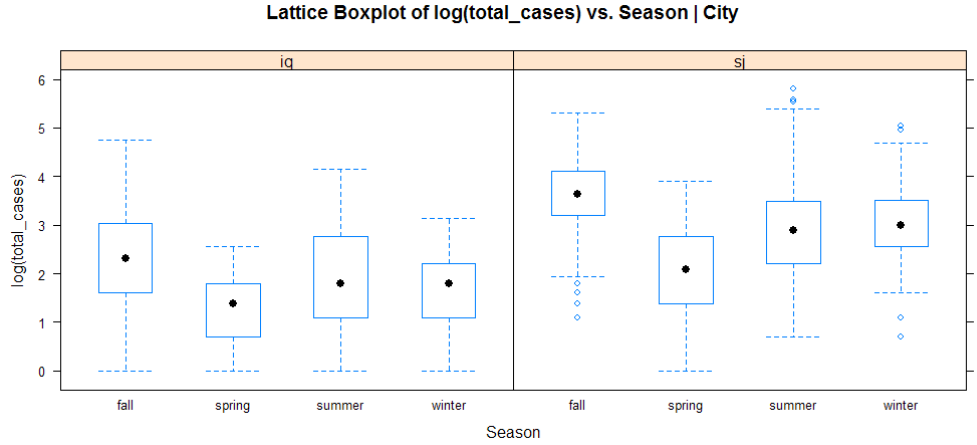


Figure 6 shows the relationship between season and log(total.cases), stratified by the city variable. On a rough look, it seems that the jumps between each season change slightly for the two cities. The Fall to Spring gap seems larger for San Juan than it does for Iquitos. In addition, the gap from Spring to Summer seems larger for San Juan. Because the relationship between log(total.cases) and season is modulated by the city variable, a modulation term between city and season will be included.

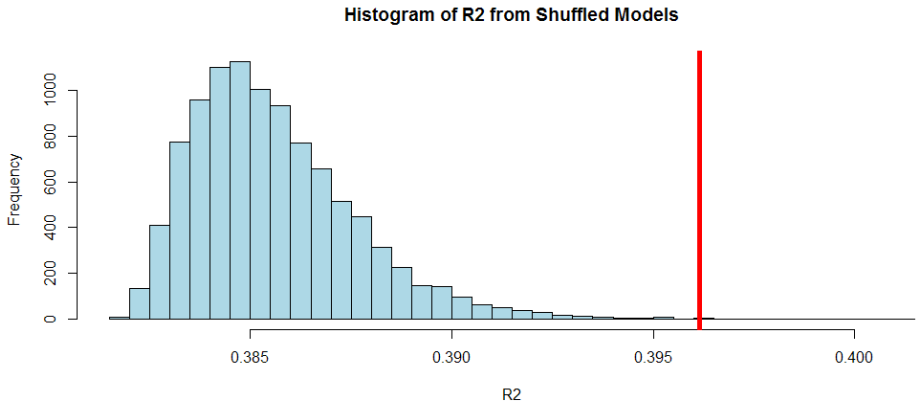
Figure 6



Because the eye test was only roughly certain, a hypothesis test was conducted to see how useful the addition of the modulation term was to the model. For ten thousand iterations, the linear regression model fit to predict log(total.cases) from the predictor variables specific humidity, city, season, and city:season was re-fit, but the modulation term shuffled. Because of this permutation, any increase in R^2 in the distribution of values is purely from randomness.

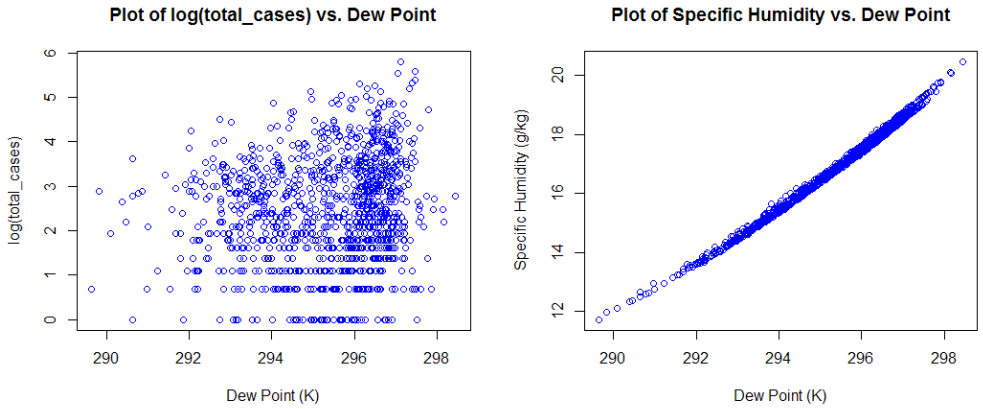
The actual increase in R^2 from the addition of the modulation term was compared to this distribution of values in Figure 7. With this distribution, it was found that the probability of getting an increase in R^2 greater than or greater than the increase actually seen given that the modulation term was not predictive at all was .04%. Because of this, it is not likely the increase in R^2 found by adding the city:season modulating term was due to pure randomness.

Figure 7



The last confounder that needs to be added to the model is dew point. No other variables in the dataset were determined to be confounders so the parsimony principle means they should be left out of this inferential mode. On the right of Figure 8 below, there is a clear relationship between specific humidity and dew point. On the left, the relationship is less certain, but has an approximately positive slope. Since dew point is correlated with both specific humidity and the response variable, it will be included in the model.

Figure 8



The final regression equation is shown below.

$$\log(total.cases)_i = \beta_o + \beta_1 specific.humidity_i + \beta_2 dew.point_i + \beta_3 san.juan_i + \sum_{k=4}^{6} \beta_k season_{ij} + \sum_{k=7}^{9} \beta_k season_{ij} san.juan_i + \varepsilon_i$$

5 Results

5.1 Parameter Estimation

	log(total.cases)		
	B	*CI*	*p*
(Intercept)	225.67	53.31 – 398.02	.010
specific.humidity	0.90	0.27 – 1.52	.005
san.juan	1.52	1.28 – 1.76	<.001

seasonspring	-0.73	-1.01 – -0.45	<.001
seasonsummer	-0.28	-0.53 – -0.04	.025
seasonwinter	-0.59	-0.84 – -0.35	<.001
dew.point	-0.81	-1.43 – -0.19	.011
san.juan:seasonspring	-0.58	-0.90 – -0.26	<.001
san.juan:seasonsummer	-0.37	-0.69 – -0.06	.019
san.juan:seasonwinter	0.30	-0.04 – 0.63	.082
Observations		1144	
R^2 / adj. R^2		40.0% / 39.5%	

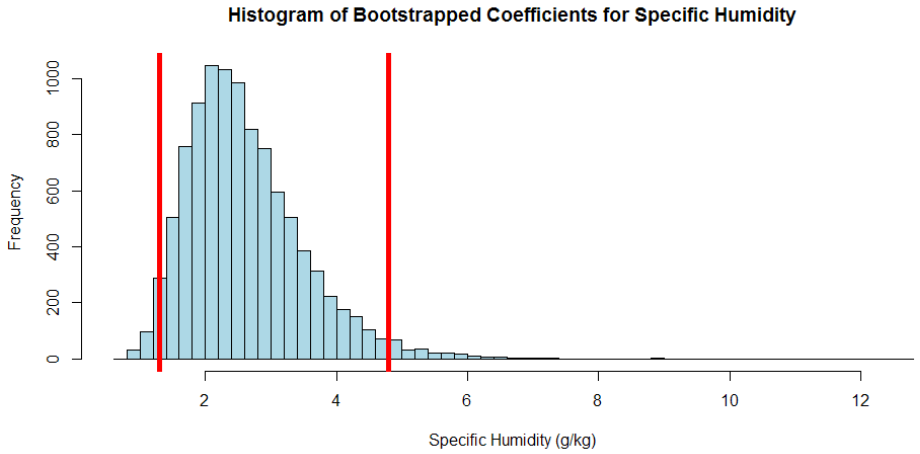
5.2 Coefficient Interpretation

Since the model is predicting log(total.cases), the interpretation of the coefficients is a multiplicative change rather than an additive change. Once the model is fit and the coefficient is exponentiated, *it is found that a one unit increase in specific humidity increases predicted cases of dengue fever by 146%.*

A bootstrapped sampling distribution, shown below, was created with ten thousand iterations to include uncertainty that arises from sampling variation. A 95% confidence interval for the coefficient was found to be (30%, 380%). Though it is a fairly wide range of plausible values, zero is not in the range so the variable is statistically significant and most values are over a 100% increase so it is practically significant as well. Increases in specific humidity are associated with huge increases in cases of dengue fever.

With an R^2 of 40%, the model is not as predictive as past research models have been, but since the goal of the model is inference, this is acceptable. As long as all counfounding relationships are included, the partial slope between log(total.cases) and specific humidity is a proxy for a casusal relationship.

Figure 9



6 Discussion

Researchers can use this paper to confirm what they currently know about the link between mosquitoes and dengue fever. The initiatives aimed at destroying the mosquito population are backed up by this model.

As previously discussed, mosquitos are the vector by which dengue fever spreads and mosquitos thrive in humid environments. This model confirms that after accounting for the effects of confounding variables like season, dew point, and city, humidity is positively correlated with dengue fever. The mechanisms by which researchers understand dengue fever are seen in real-life data.

In San Juan and Iquitos, specific humidity ranges from 12 to 20g/kg. On a week where specific humidity is only 3 to 4 points higher than expected, this model suggests that dengue fever cases could be 14x to 34x higher than usual. More than being statistically significant, this finding is practically significant. Orders of magnitude that large are enough to separate out the smallest outbreaks of dengue fever out from the largest outbreaks.

As compared to past studies on this topic, the model developed by this paper has a few differences. Compared to Sintorini's 2018 Jakarta study, this model takes into account confounding variables. Compared to McMurren's 2013 Paraguay study, this model focuses on inference rather than predictive accuracy.

By seeing similar results in another two cities, the assurance that health officials have that this is a useful predictor goes up. Flukes can happen in a few cities, but as more papers find the same effect of humidity on dengue fever, the odds of a giant fluke go down. World health officicials should use this information to expand their efforts of mitigating mosquito contact in

humid areas. Programs to remove still water where mosquiots breed and give out protective clothing to people will likely be effective.

7 Conclusion

There is a need to allocate scarce resources among many possible cities to fight dengue fever. Since mosquitoes spread dengue fever and thrive in humid environments, an inferential, multiple linear regression model was built to estimate the partial slope of the relationship between specific humidity and total cases. The partial slope was estimated to have a multiplicative effect of a 146% increase in predicted cases of dengue fever for every one unit increase in specific humidity (g/kg). The more humid an area is, the more cases of dengue fever are expected. With this model and past research together, dengue fever cases can both be better understood and planned for.

8 References

Boussion M. (2012). "Predicting outbreaks of dengue fever according to climate".

Fatima, Q. (2018, May 28). "The other side of Big Data - Dengue Fever Prediction Models".

Guzman, M. G. et al. (2010). Dengue: A continuing global threat. Nature Reviews Microbiology 8, S7–S16

Henchal EA, Putnak JR. (October 1990). "The dengue viruses". Clinical Microbiology Reviews. 3 (4): 376–96.

McMurren J, Young A, Verhulst S. (2013). "Forecasting outbreaks with open data".

Sintorini M M. (2018) IOP Conf. Ser.: Earth Environ. Sci. 106 012033

WHO. (May 2015). "Dengue and severe dengue Fact sheet Nfl117".

Appendix

Appendix A: All dataset variables and explanations

total.cases – Total recorded number of dengue fever cases for a week in one of two cities.

city – City the data was recorded.

season – Season the data was recorded.

ndvi.ne – This variable is the NDVI measure for the northeast part of the city. Normalized Difference Vegetation Index (NDVI) is a measure of how much vegetation surrounds an area. Determined through satellite imaging, the metric tends positive for areas of high vegetation like a rainforest and negative for sparse areas like a snow field.

ndvi.nw – NDVI for the northwest part of the city

ndvi.se – NDVI for the southeast part of the city

ndvi.sw – NDVI for the southwest part of the city

precipitation.amt.mm – Rainfall for the week in millimeters

air.temp.k – Average air temperature in Kelvin for the week

avg.temp.k – Average ground temperature in Kelvin for the week

dew.point.temp.k – Average dew point in Kelvin for the week. A dew point is the temperature to which air must be cooled to become saturated with water vapor.

max.air.temp.k – The maximum recorded air temperature in Kelvin for the week

min.air.temp.k – The minimum recorded air temperature in Kelvin for the week

precip.amt.kg.per.m2 – Rainfall for the week in kilograms per meter^2. This is a more relative measure of precipitation compared to raw precipitation in millimeters.

relative.humidity.percent – Average relative humidity in percent for the week. Relative humidity is a measure of how saturated the air is with water. 100% humidity is completely saturated air. This is a relative measure of humidity that is dependent upon what the temperature is.

specific.humidity.g.per.kg – Average specific humidity in grams of water per kilogram of air for the week. This is a raw measure of humidity based purely on how much water is in the air.

tdtr.k – Average Diurnal Temperature Variation (DTR) for the week. DTR is the difference between the maximum and minimum temperature for a single day.