Opening a Cold Case

Investigating Temperature's Effect on The Rate of Robberies

Aaron Maurer

December 4, 2014

1 Introduction and Objective

There have been a number of studies on how temperature affects the rate of violent crimes, in particular assault. The consensus seems to be that rate of these crimes rises in hot weather, with the inclination being that it also drops off in cold weather. Both these changes are thought to occur due to the effect of temperature on human emotion[1]. However, I was curious what relationship exists between a more practical and presumably less emotional crime such as robbery and the temperature.

To investigate this question I have endeavored to quantify how the rate at which robberies occur in the city of Chicago varies with temperature. Robbery, defined as taking property from a person, without their consent, by force or threat of force[2], includes what is probably the most common outdoor crime, mugging. It would seem natural then that the rate at which these occurred would be sensitive to the weather. As well, investigating this relationship will give insight into when one is at the greatest risk for being robbed should they go out, though the number of people out and about also factors in.

2 Data

The Chicago city government publishes a data set which includes every crime reported to the police in the city, going back to 2001[3]. I made use of the observations from this data which had a date recorded for the crime during 2011, 2012, or 2013, and which had a primary description of "Robbery", per the Illinois Uniform Crime Reporting Code. These crimes include both successful and attempted armed robbery, unarmed robbery, and vehicular hijacking. For each crime, the data set has additional information on where it occurred in the city, more detailed information about the crime, whether the crime was domestic, and if it resulted in an arrest. I, however, simply aggregated the total number of reported robberies that occurred anywhere in the city during each hour of each day during the time period.

I combined this hourly data with top of the hour weather data from O'Hare International Airport[4]. This data is a series of meteorological data, including temperature, accumulated precipitation, humidity, and wind speed, almost always measured once during a particular hour at 51 minutes after that hour. When there were multiple entries for an hour, I chose the first at 51 minutes which had the temperature recorded, or the first chronologically which didn't have temperature missing, discarding the rest. This was taken as the temperature for the city for a particular day and hour, and was matched with the hourly robbery counts. The small number (≈ 10) of hours for which there was no recorded temperature from O'Hare were excluded. The end result was $26,263$ hourly observations of temperature and reported robbery count, covering almost all of 2011-2013[5].

[1] Andrea Anderson, "Links Between Crime & Chilly Weather is Complex, Controversial", *Scientific American*, January 1st, 2014, Accessed December 4th, 2014. http://www.huffingtonpost.com/2014/01/01/weather-crime-temperatures_n_4518940.html

[2] As opposed to theft, which is the taking property without consent but not necessarily by violent means, and burglary, which is breaking into a structure so as to commit a crime (whether or not a crime such as theft is committed)

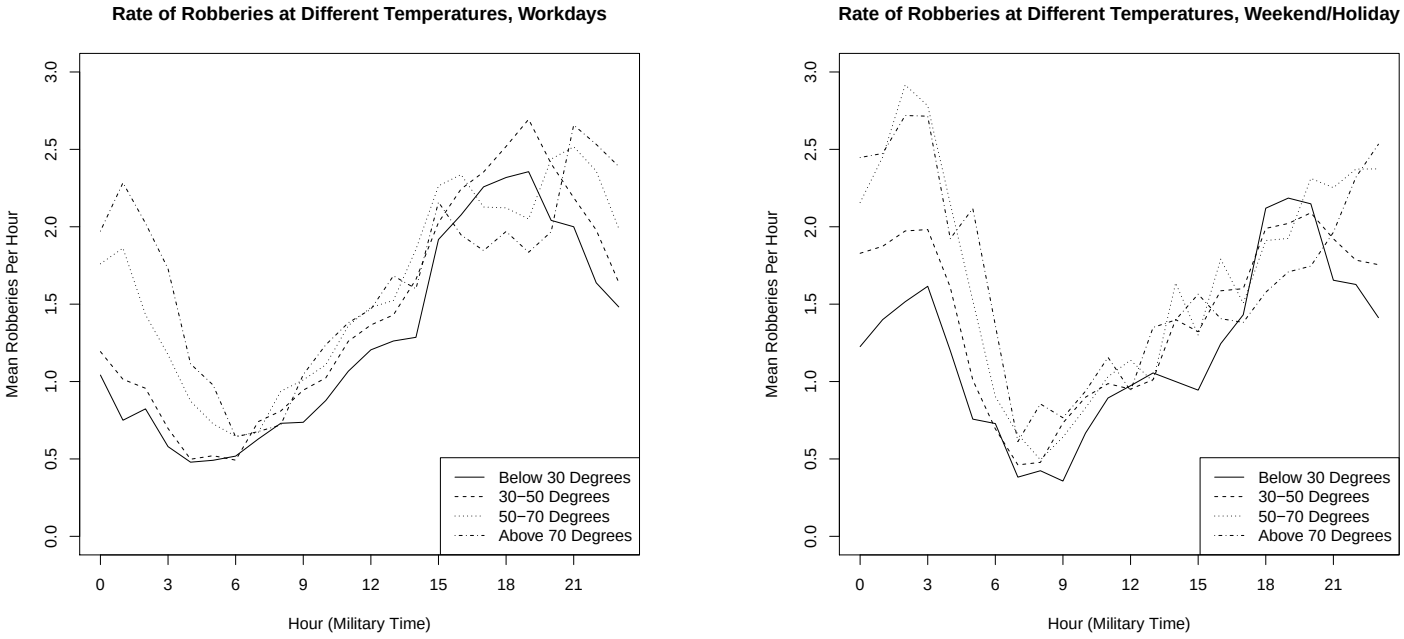
[3] City of Chicago. (2014). *Crimes - 2001 to present.* Retrieved from https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

[4] Midwest Regional Climate Center. (2014). *Unedited Hourly Data - Top of the Hour Observations - O'Hare International Airport.* Retrieved from mrcc.isws.illinois.edu/CLIMATE/ucld/ucld_hrlyTop_getdata1.jsp?WBAN=94846

[5] My R script is available on github: https://github.com/ajmaurer/Chicago-Course-Work/blob/master/stat374/final_project.R

Examining this data, it is obvious that the most important feature in determining the number of robberies is the time of day, and to a lesser extent, time of week. Figure 1 speaks to this effect, displaying an average by hour of the day for different temperature ranges and workday versus holiday/weekend. The mean number of robberies peaked between 6pm and 3am, occasionally twice in that span. The minimum was consistently around 6am, with the mean at the peak 4 to 5 times as high as at the trough. This corresponds roughly to typical people's schedules; there were the most robberies during the period where people have gotten off of work for the evening, and the least when the fewest people are awake early in the morning. During the weekend, the trough and peak tended to be later, as people were up later or sleep in.

Figure 1



Note: With fewer observations, the Weekend/Holiday series can be expected to be noisier

The effect of temperature seems to be smaller than this, and also varies with the time of day. During the middle of the day (8am to 3pm or so), the means were similar for all temperatures. However, there were fewer robberies in the late evening and early morning when it was cold. One of the most interesting features is that in the evening, there were fewer robberies between 6pm and 7pm when it was warm, with a much higher rate before and after. This is likely at least partially due to the changes in people's schedules that occurs around summer, when it is both warmer and light later.

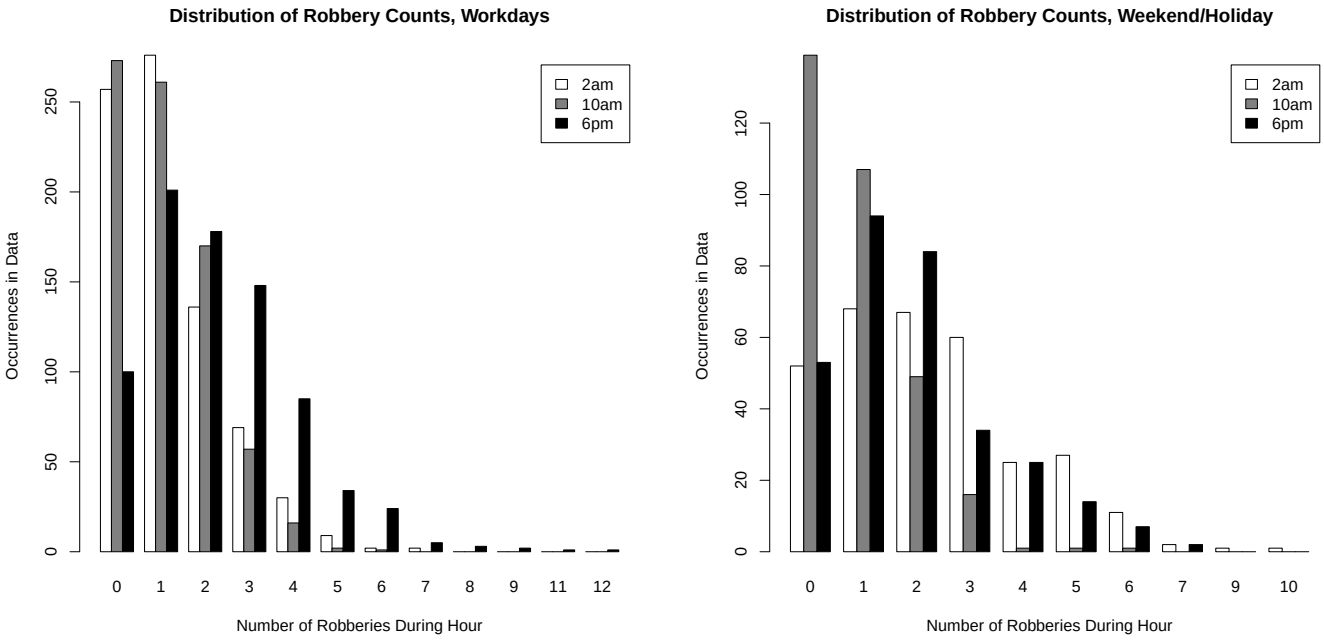
Looking at the distribution for a sample of hours in Figure 2, we once again see the effect of time of day and workday versus weekend/holiday. However, we also see a data set which seems to have, as one would hope for with count data, a distribution which is approximately Poisson.

3 Parametric Model

This observation points to the obvious parametric model to fit to the data, a Poisson regression model[6]. This is a natural choice, and considered the first parametric model to try for count data, if not the default. Its assumptions are several fold. First, it assumes that data is Poisson distributed around a particular mean, which arises when counting the sum of infrequent and independent occurrences of which a large number of these events could, in theory, occur. This assumption seems fairly safe; robberies are infrequent, and though one criminal could be responsible for multiple robberies, this effect overall can be expected to be trivial in such a large city. The less safe assumption is that the mean of this distribution can be predicted by the product of exponential functions of the set of predictors. It isn't obvious that this is true, but it is mathematically convenient, constricting the mean to positive numbers.

[6] A generalized linear model with a Poisson distribution and log link fit by maximum likelihood

Figure 2



Choosing this as my model, I tried a number of specifications on an 80% training subset of the data. These included various interactions of hour, as a factor and bucketed, temperature as a continuous variable, and workday versus weekend/holiday as a factor. Choosing the specification that had the best trade off of number of parameters versus likelihood, as dictated by the Akaike Information Criteria, I settled on an specification that included as predictors:

- Hour of the day as a factor
- Buckets of three hour periods as a factor interacted with workday as a factor
- Buckets of three hour periods as a factor interacted with the log of the sum 50 plus temperature.

Implicit in this set up is that the effect of temperature within a particular hour is assumed to be monotone.

The estimates of the coefficients themselves are not particularly elucidating; the overall effect is best displayed graphically, such as through the heat maps in Figure 3. Here, we see the model capturing several of the features we noticed earlier. The predicted robbery rate is quite low in the late evening and early morning when it is cold, and rises quickly with the temperature, especially on the weekend. A bit later (6am or so), the crime rate is low, and throughout the middle of the day it only rises slightly with temperature. The early evening is the only area that seems outright strange; high values are predicted at low temperatures, and drops off somewhat as the temperature increases. This is questionable, but possibly an artifact of the assumed monotone relationship between temperature and robbery rate paired with the small dip between higher and lower temperatures in this range.

The effect of temperature is displayed again individually for 2am, 10am, and 6pm in Figure 4 along with 95% confidence intervals. Here, we see again the patterns listed above, but also that the model predicts these big swings in robbery rate due to temperature with a great deal of certainty. Particularly, at 2am, the confidence intervals for two predictions with as little as 20 degrees separation don't overlap, asserting that they are different with much higher confidence than 95%.

Figure 3

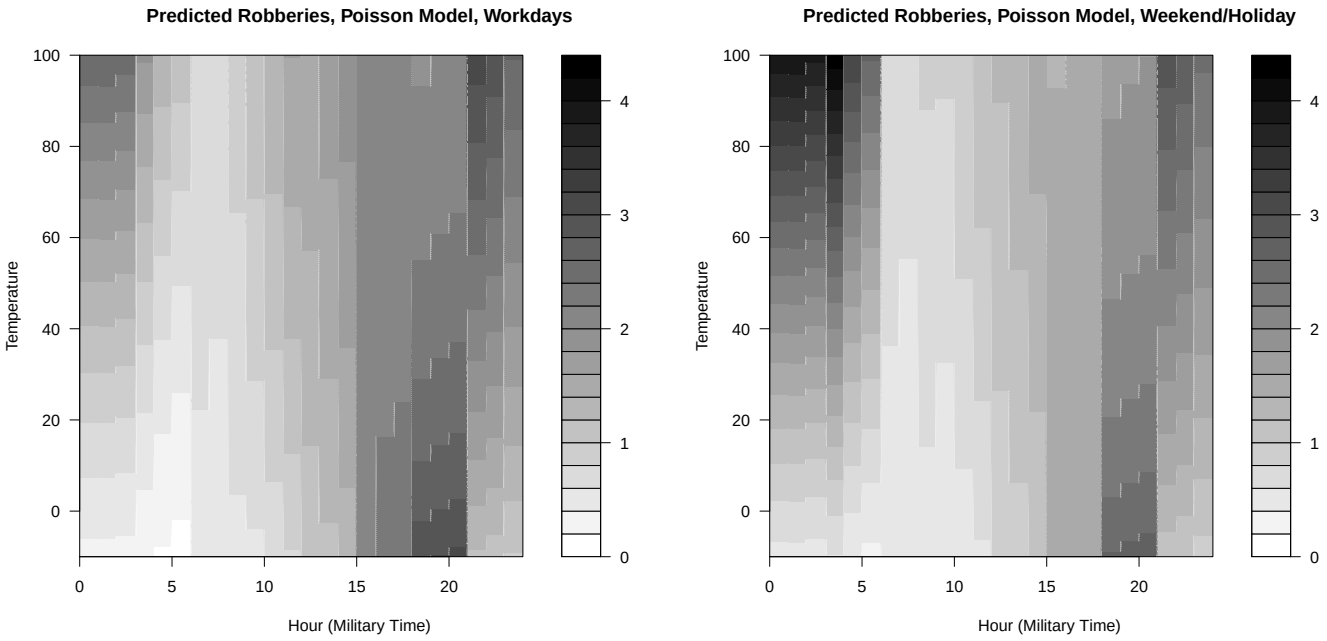
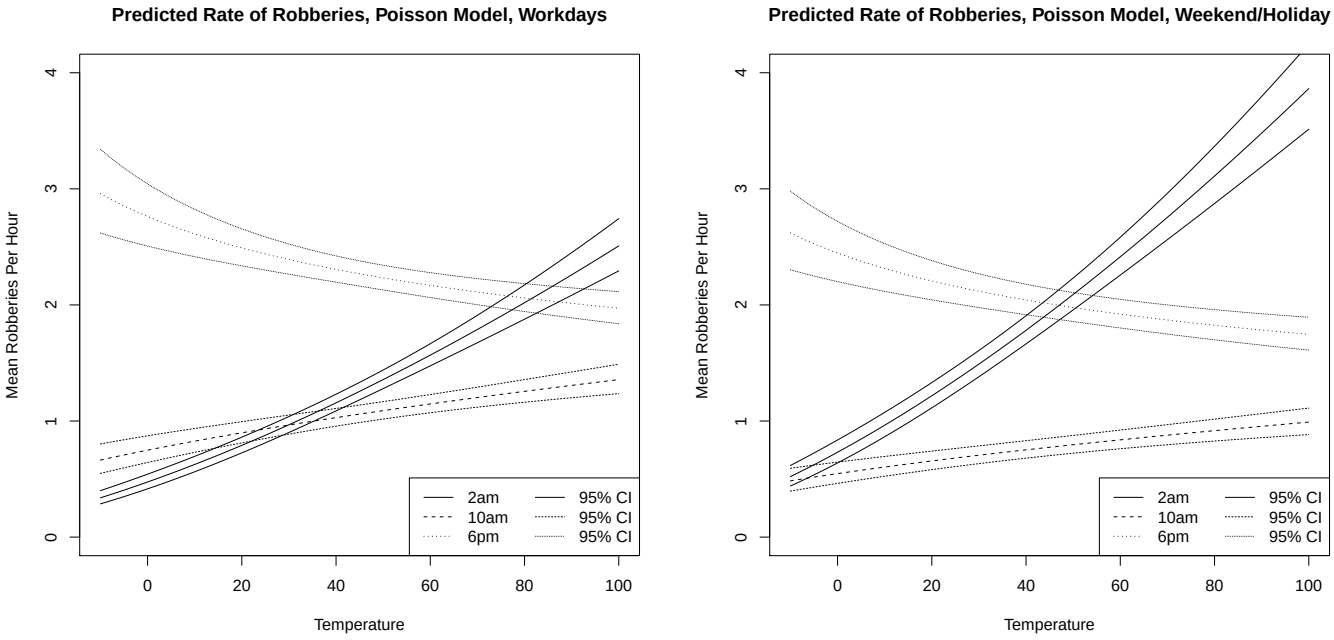


Figure 4



The main danger with a Poisson regression model, such as this, is that if the predicted mean function may not be able to capture fully how the mean varies. This would most likely be due to insufficient predictors being included. The result may be larger variance around the predicted mean than allowed for with the Poisson Distribution, invalidating the model. In this case, should the rest of the assumptions hold, a negative binomial regression can be used, which can be derived as a certain sum of Poisson distributions with varying means, as could be the case should the mean not be perfectly predicted by the predictors. This makes the negative binomial model a generilization of the Poisson model, and a formal test can be used to decide if the additional parameter of a negative binomial is needed. Such a test might well have indicated that I should have used a negative binomial regression here, but its unlikely the resulting predictions would have been wildly different.

Additionally, as with nonparametric models, a better tool for evaluating the fit of a particular specification of the model may have been cross validation rather than Akaike Information Criteria, since it gives a more direct

estimate of the suitability of the fit.

4 Nonparametric Model

To model this data with a nonparametric model, I ran separate local linear regressions on workdays and non-workdays, with each predicting the log of the sum one plus number of robberies in terms of hour and temperature simultaneously. As with the Poisson model, I fit on an 80% training subset. In fitting the model, I used a Gaussian kernel, individual bandwidths for each point based on a global nearest neighbor fraction[7], and a relative scaling factor for the second variable[8]. These two values were chosen from a grid of values based on which resulting model had the lowest leave one out cross validation score. As well, the hour span was expanded to include the previous and next day's data, so there should be no boundary bias towards the beginning and end of the day.

The reasoning and justification for these choices are as such:

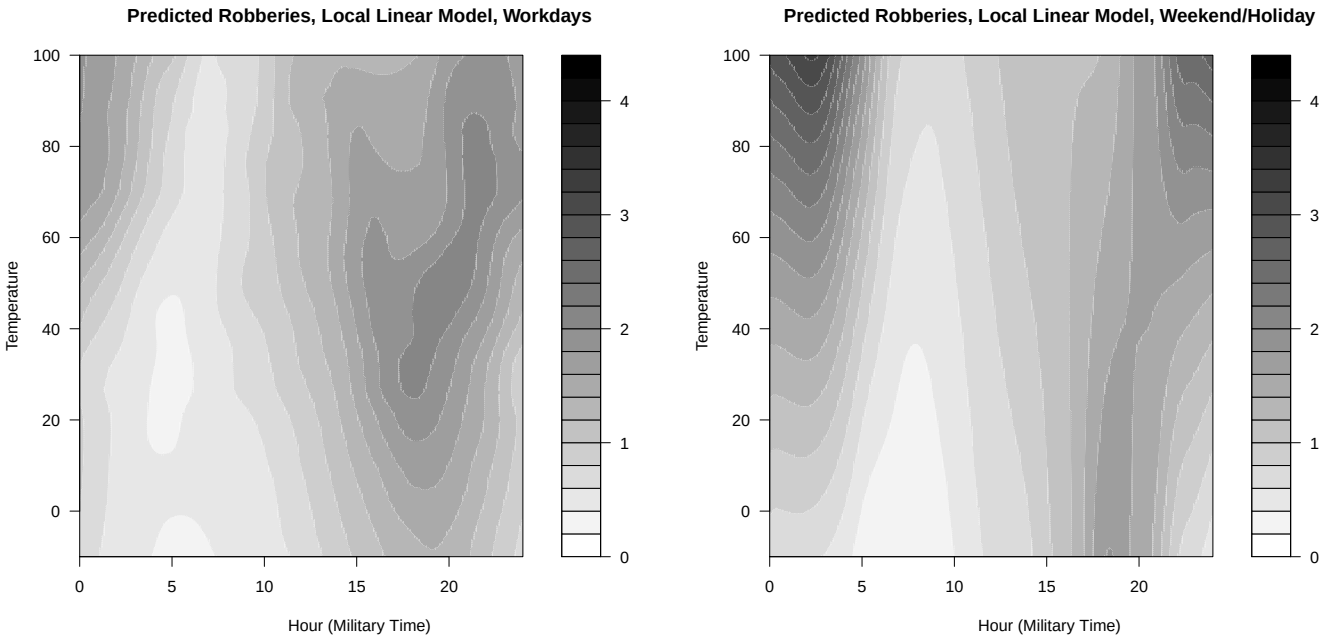
- The kernel doesn't make a great deal of difference, so the choice was arbitrary.
- To study the effects of temperature at the extreme end of their range, I need a model with low boundary bias, making a local linear fit superior to a simpler kernel regression. Higher polynomial fits on the other hand make it increasingly difficult to fit the model well.
- Predicting log one plus robberies nicely constrains the prediction to positive counts, which is what it should be in the real world.
- As the initial plots have shown, there seem to be fairly distinct trends between workdays and not workdays, so two separate models seemed more appropriate than one.
- The relationship between temperature and robberies appears to vary by time of day, which would not have been captured by a generalized additive model which modeled the two variables without interaction. Since there is sufficient data to achieve satisfactory narrow bandwidth even modeling the two variables together with one local linear model, this choice was preferable.
- The observations are not equally spaced out over their range, so one bandwidth for all the points wouldn't smooth enough on the extremes or too much in the middle. Instead, by using a nearest neighbor fraction, the amount of smoothing was comparable across the data points.
- Including a scaling factor for the second variable is essential, since the units of the two variables aren't comparable.

With this model, two assumptions are being made. First, the number or robberies as a one time and temperature is correlated with the number of robberies at similar times and temperatures, which is obvious. Second, that there sufficient smoothness in the underlying function to achieve convergence of the estimates. Having looked at simpler plots, this also seems likely.

[7]The bandwidth is set for each point such that the fraction of the total data less than the bandwidth's distance from the point is the desired nearest neighbor fraction

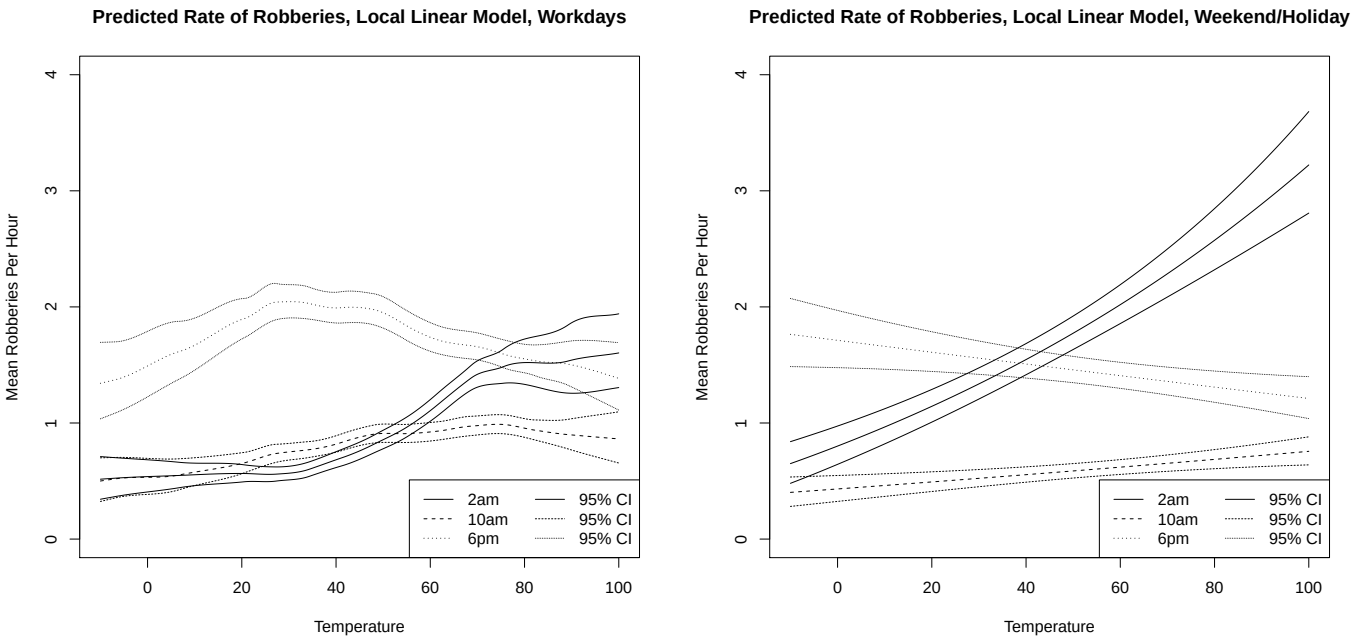
[8]This factor is used to scale the second variable with respect to the first for determining the distance between two points for the weighting function

Figure 5



The predictions of the models can be seen in the contour plots of Figure 5. It is important to note that with less data, bandwidths for the weekend/holiday model were chosen to include a larger portion of the data around each point, and also that the scale was such that this model averaged more strongly over temperature than hour as compared to the workday model. This model seems to very clearly portray the patterns we have seen before. Peak crime on workdays is highest in the evening, with the main peak moving later and a second smaller, earlier peak developing with hot temperatures. Also, the afternoon and evening crime rate does fall off as the temperature increases beyond a point. On non-workdays, the peak is in the earliest hours of the morning, and keeps increasing with the temperature. However, the prediction for the highest temperatures in the early morning should be taken with a grain of salt; their is little if any data where temperatures actually reach that high at that hour of the day.

Figure 6



Evaluating the confidence of the fits, as seen in Figure 6, we see a much more nuanced picture for workdays than we saw from the Poisson model. It should be noted that the confidence intervals are based on local estimates

of variance, necessitated by the variable bandwidth. This model is quite certain that the peak crime rate at 6pm on workdays occurs when the temperature is between 30 and 50 degrees, and falls off on either side, as opposed to the monotonic relationship predicted by the Poisson model. It also has a large amount of confidence there is a true increase in crime in the early morning, though as I noted, at high temperatures there is much more uncertainty. With much more smoothing, particularly over temperature, in the weekend/holiday model, its results look similar to the Poisson models.

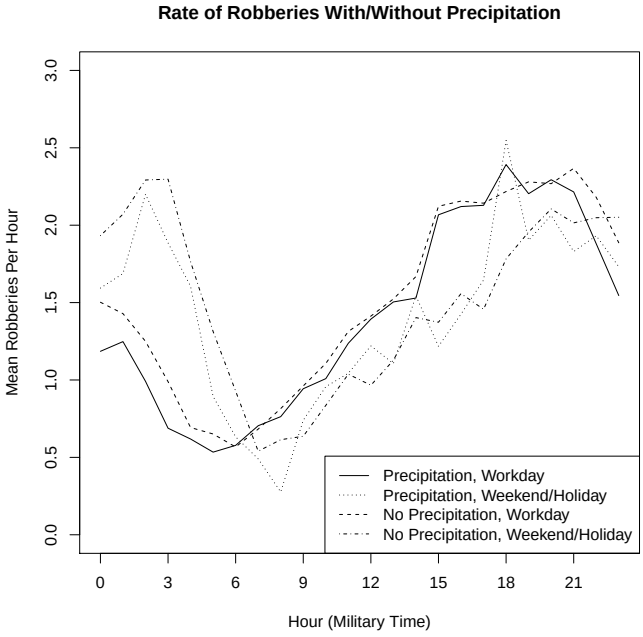
It seems safe to say this model is the strongest non-parametric regression for this subject. Alternate linear smoothers and generalized additive models were discussed earlier. Monotonicity with hour or temperature doesn't seem to hold, nor does convexity/concavity, so models making those assumptions aren't appropriate. This series doesn't have the quickly varying features that would necessitate a procedure like wavelet regression. However, what could be quite a good alternate approach is to not have aggregated robbery counts, and fit a density estimate to the data. This could be scaled by how often different temperature and hour combinations occurred to give relative rates of robbery at a given temperature and time (per the definition of conditional probability). This approach has the advantage of preserving more information about the time a crime occurs, but is computationally more intensive (with more data points), and doesn't easily admit a comparable parametric model.

5 Discussion

Comparing the two models just based on performance, the nonparametric model wins, though they both do fairly well. I calculated the root mean squared log error for each model over the 20% test subset that wasn't used to train the models. The local linear model had a RMSLE of .510, while the Poisson model had a RMSLE of .529. By comparison, an unbiased estimate where the mean at each temperature/hour combination in the training data was taken had a RMSLE of .563, so both models are a meaningful improvement.

Subjectively, they both picked up mostly the same trends. The only notable distinction was in how they treated the late afternoon and early evening period, with the Poisson model apparently failing to capture the non-monotone trend. However, by including more and varied predictors (from interactions, splines, transformations, etc.), one can get and arbitrary degree of flexibility. Choosing a sufficient set of predictors is a slow process to do manually though, making it easier to just start with a non-parametric method which has arbitrary flexibility built in. One could meet halfway between the methods with a nonparametric model fit with local likelihood, since the assumption that the error is Poisson or negative binomial distributed seems appropriate.

Figure 7



In terms of interpretation, both fairly convincingly show that the rate of robberies varies with temperature,

generally increasing with it. Causation is quite unclear though; temperature is highly correlated with a number of other seasonal factors; one could reasonably expect the rate of robberies to vary with school holidays, daylight hours, or just the time of the year irrespective of season. By including all these factors in a model, it would likely be highly predictive, but the high collinearity can be expected to make it hard to attribute changes to just one factor, modeled parametrically or not. As well, though three years is a relatively short period for policing and social changes, there is a well known secular decreasing trend of urban crime in America which was not accounted for. Should this have synced with the relative severity of winters, it may have skewed results. Appropriately modeling these effects would be the most important extension of the current model.

There are a number of other areas where this model could be expanded though. Crime rates are quite heterogeneous throughout the city, in a fashion that may vary with weather, and there is a great wealth of geographic data in the Chicago crime database to model this. Different types of crime could also be investigated. Different attributes of the weather may also play a roll; I initially investigated the effect of precipitation on the robbery rate. this seemed to show that presence of precipitation, as seen in Figure 7, does seem to coincide with a small reduction in robberies, at least in during the late evening and early morning. This effect appears to vary a lot with the amount of precipitation, and there was relatively little data from hours with enough precipitation to have a meaningful effect. With the Poisson model, every version I tried which included precipitation in some form had an inferior AIC, so ultimately I decided not to include it, at least for now. Finally, there is slight asymmetry between the early morning and late evening predictions due to how the weekends are defined. Friday evening is likely more akin to a Saturday night, with more robberies, pushing up the late evening average of workdays as compared to the early mornings. Sunday evening likely has the opposite effect on the Weekends/Holidays. A modified delineation between the two periods may have done away with this, and there may be further clustering of hours/days within the week that could improve the fit. These are all areas which present opportunity for further refinement.