A Statistical Analysis of Chicago Traffic

Tristan Rasmussen

December 4, 2014

1 Introduction

Congested roadways, creeping progress, bumper to bumper gridlock; these are all hallmarks of the largest cities in the modern world. Everybody hates traffic and some recent studies have shown that it is costing Americans billions of dollars every year. In this paper we aim to take the first steps in addressing the problem of traffic by performing statistical analysis on historical traffic data from the city of Chicago. We have released the code used to create these models,[1] and constructed an interactive web application to explore the resulting models.[2]

1.1 Goals

Our ultimate goal in this paper is to build models that offer predictive and explanatory power for traffic Chicago. To do so we will explore two statistical problems. The first problem that we will examine is that of estimating the probability density of traffic speed; this will help us gain a more intuitive understanding of speed and can be used in future analysis. The second problem we will explore is how we can build a regression model for traffic speed; if the models provide a good fit these will allow us to explain past traffic trends and possibly predict future traffic trends. We will apply both parametric and non-parametric methods to both questions, and perform a comparison at the end of the paper.

2 Description of the Data

In this paper we made use of two data sets: one data set containing information about traffic speeds in Chicago over a two year period, and one containing data about sports games in Chicago over the same period.

2.1 Chicago Congestion

The main data set of interest is a record of estimated traffic speeds in 29 regions of Chicago gathered from 2011-03-12 - 2013-03-26 [1]. For each data point, CTA busses were polled for their speeds and these speeds were averaged to produce a regional speed. Data was collected every ten minutes over the entire time period. In addition to the time of measurement and the average speed, each data point was supplemented by the number of buses measured, and the total number of readings used.

It is somewhat difficult to visualize the entire data set given the long time span and frequent data reads. However, we can still get a sense for some of the structure in this data by examining speeds is specific regions as a function of some of the dependent variables. For example, Figure 1 shows speed vs hour in region 13 (the downtown loop area); one can clearly see a dip during the working hours, especially around 5 PM (most plots presented in this paper will be speed vs hour, as this shows interesting behaviour without presenting so much data that it is impossible to interpret). To help the reader get a sense for the regions of interest, Figure 2 shows the 29 regions from the data set overlaid on a map of Chicago.

[1] https://github.com/courageousillumination/traffic-analyzer

[2] http://chicago-traffic.mooo.com

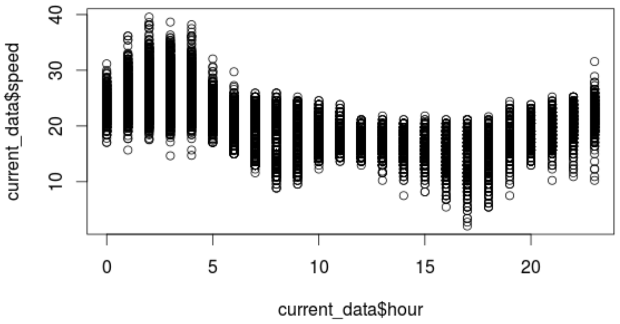
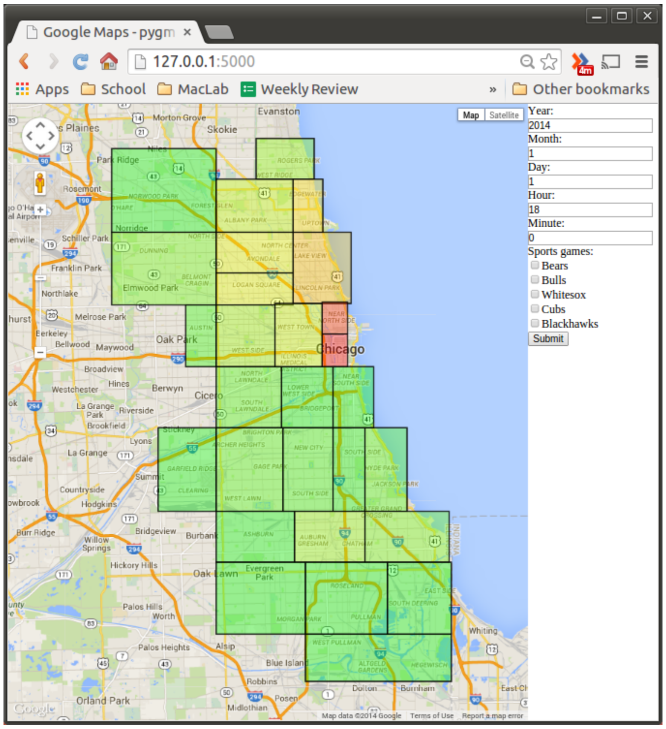


Figure 1: Speed vs Hour in Region 13



Year:
2014
Month:
1
Day:
1
Hour:
18
Minute:
0
Sports games:
□ Bears
□ Bulls
□ Whitesox
□ Cubs
□ Blackhawks
Submit

Figure 2: Regions from Chicago Congestion dataset (color represents average speed)

2.2 Chicago Sports Games

To supplement the speed data set described above we also collected a list of all sports games that happened in Chicago during the period of interest. Our goal was to identify some additional factors (other than time) that would have a significant impact on traffic speeds. Unfortunately, we were unable to find a single comprehensive list of all sports games in Chicago and were forced to compile a full data set from various sources. To do this we gathered the dates of every Bulls [2], Cubs [3], Bears [4], Whitesox [5], and Blackhawks [6] game in the time period and created a data set where each point was a date and the games that occurred on that date.

To see the effects of sports games on traffic speed we separated the speed measurements in region 11 (which contains the United Center, the home of both the Bulls and the Blackhawks) into days where there was a Bulls/Blackhawks game and all other days. Figure 3 shows the speed on game days and non games as a function of the hour. We can see a significant drop in speed on game days in the later evening; this is another feature that we would like our model to capture.

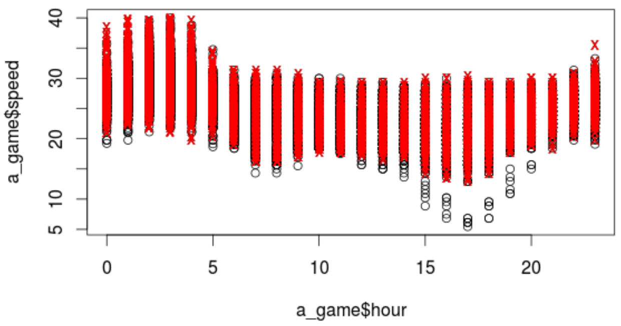


Figure 3: Speed vs Hour in Region 11 with (black Os) and without sports games (red Xs)

2.3 Data Preprocessing

To produce our final data set we started by combining the two data sets above. Each data point of the traffic speed data was augmented by adding in all the sports games that occurred on that day. Additionally, to make regression easier, we broke down the timestamp for each row into a sequential day label, year, month, day, hour, minute, day of week, and a boolean flag that indicated if the day was a weekend. Finally, we added a column that indicated if there was a sports game of any kind.

Having created our final data set, we performed some preprocessing to deal with outliers.

First we removed all the data points that recorded messages from zero buses as these are clearly invalid reads. Next, we removed all rows that recorded a speed over 60 MPH; all the roads in consideration had speed limits well under 60 MPH, and we considered any reads higher than this to be anomalous. Finally, we removed all rows that recorded a speed of 0 MPH; although this is a possible valid value, it occurred much more frequently than could realistically be expected and we were unable to discern valid reads vs invalid reads. Finally, a small amount of noise was added (drawn from $N(0, 0.01^2)$) to deal with repeated values; this is significantly smaller than the noise present before hand and shouldn't significantly affect the results.

3 Parametric Analysis

In this section we discuss the parametric methods we applied to the traffic dataset. All of the numeric results from this section are collected in Table 1. Both density estimation and regression were run over all 29 regions separately; for the sake of brevity we only include plots for individual regions or discuss them if they show interesting behaviours or are exemplary of a broader class of regions.

3.1 Density Estimation

To create a parametric estimate of the density of traffic speed it was necessary for us to assume a single model for the underlying distribution. We chose to model the density as a single normal distribution. We had originally considered modelling the data as a mixture of Gaussians, but we ruled this out since preliminary analysis showed the data to be mainly unimodal, and lacked the wide peak that is indicative of a mixture of Gaussians. We also considered modelling the data using other well known distributions but none of these seemed to fit the data particularly well.

Once we assumed this model of the data, we were left with the problem of parameter estimation. We chose to use the maximum likelihood estimators for estimating the parameters. As an alternative we could have employed a Bayseian method such as maximum a posterior estimates, but for this paper we wanted to approach this data from a frequentist perspective.

Given this model and method for estimating perameters we could fit our density by simply calculating the sample mean and standard deviation and using these to construct a normal distribution. We recorded the estimated mean and standard deviation in Table 1 along with 95% confidence intervals for both. We also produced plots of each fitted density; however, these plots are more useful when compared to the nonparametric density estimation and so will be presented in the Nonparametric Density Estimation section below.

In addition to estimating the distribution for speed in each region we generated QQ plots to try and get a sense of how well the data fit our model. Unfortunately, these plots showed significant nonlinearities, indicating that the distributions deviated from our normal assumption. Moreover, the deviations were not consistent across regions and generally didn't match any standard distribution that we were aware of. For example in region 13 (Figure 4) it is mostly linear until the far right side at which point is experiences a significant spike; in contrast region 16 (Figure 5) shows many nonlinearities especially on the left side of the plot. Thus, although it became apparent that our normal assumptions were not necessarily valid, we were unsure how to proceed in a parametric manner.

3.2 Regression Analysis

Having looked at density estimation we then turned to training parametric regression models for traffic speed. In the course of our analysis we trained and tested two different classes of parametric models: linear models, and cubic models. Both of these models make the assumption that the data come from some linear/cubic function of the data plus some additional noise that is drawn iid from a zero mean normal distribution (in reality, this assumption may be false; see Future Work). We wanted to form linear model as something of a

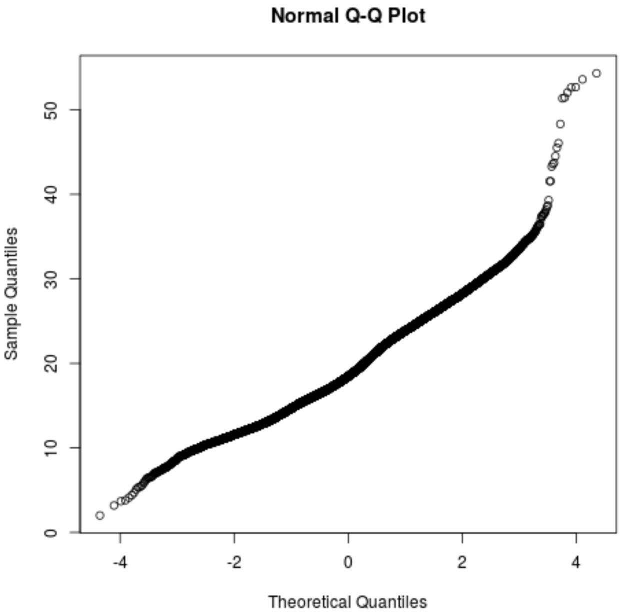


Figure 4: Region 13 QQ Plot

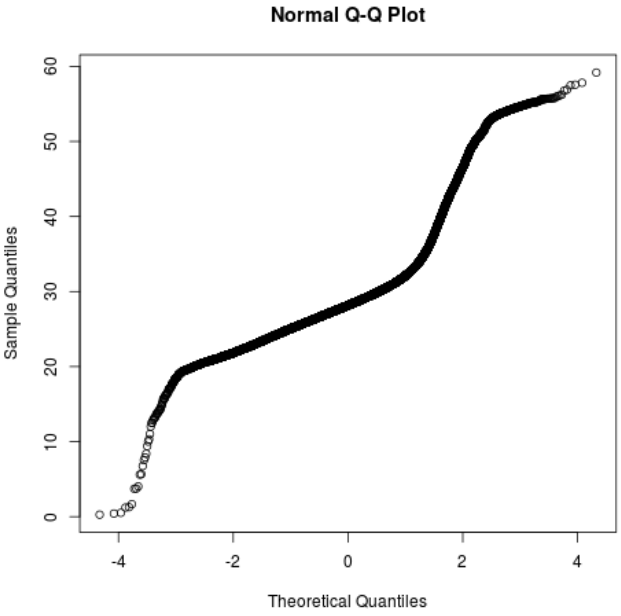


Figure 5: Region 16 QQ Plot

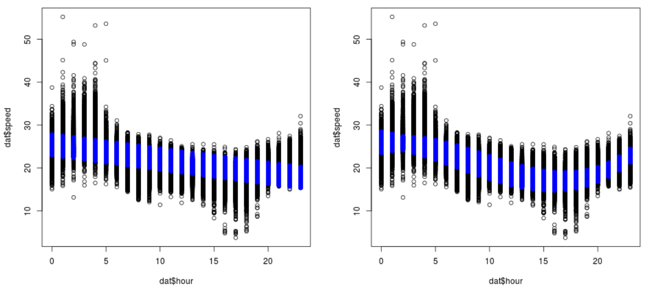
baseline; these being one of the simplest possible models would give us a sense of how easily we could fit the data. We opted to form cubic models under the assumption that linear models would be insufficient to fully explain the data, and that more complexity would be needed. We chose cubic models to avoid over fitting after some exploratory analysis indicated that the majority of the response was cubic in the input. We chose to fit these models using least squares regression, mainly for its computational speed and ease of use.

In both the linear and the cubic cases we first partitioned the data into training and test sets; the test set was a 25% random sample of the whole data and the training set was the remaining 75% (we added the ability to seed the random number generator to allow for reproducible runs). After we partitioned the data we performed model selection using Akaike information criterion (AIC). Due to time and computation constraints we were unable to examine all possible models; instead we hand selected a set of models that we expected to fit the data well. We then trained each of these models on the training set and chose the model with the lowest AIC. Finally we ran the best model against the test set and computed the testing error. Table 1 contains the training error, test error, and r^2 for the best linear and nonlinear models. We also computed confidence interval for the coefficients in each regression, but due to space constraints they will not be included in this paper.

From this analysis we can see that the results varied greatly between regions. For example, the regressions performed very well in region 12 (figure 6) where we achived a r^2 of 0.64 for the cubic model. However, the fit was much worse in region 28 (figure 7) where the r^2 from both fits was less that 0.1. We will explore why we see these vastly different results in the Discussion section.

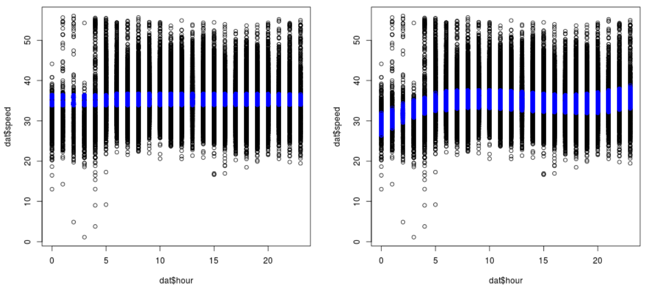
4 Nonparametric Analysis

In this section we wanted to analyze the dataset nonparametrically, making much weaker assumptions than we did above. Once again we created models for each region separately and all numeric results are summarized in Table 1.



(a) (b)

Figure 6: Regression and real data vs hour for region 12



(a) (b)

Figure 7: Regression and real data vs hour for region 28

4.1 Density Estimation

We saw above that parametric density estimation hit a dead end after producing QQ plots and realizing that the normal assumption was invalid; it seems that the data doesn't conform well to any well known distribution. As such this is a perfect use for a nonparametric density estimate. In our case we chose to use kernel density estimators (KDE). This method makes very weak assumptions on the underlying density (mainly that the true distribution has two continuous derivatives), and has a theoretically optimal rate of convergence.

Our KDEs were created using an Epanechnikov kernel, with bandwidth chose using Silverman's rule of thumb. We then plotted each fit along side the parametric density estimate.

Comparing the KDE to the Gaussian estimate, we can see that the KDE shows significantly more detail of the underlying distribution, without producing excessive irregularities. For example, consider region 13 (Figure 8); under the parametric model we discern any interesting features, but using the KDE we can see a small, but definitely present, hump on the right side of the distribution. In other regions, such as region 22 (Figure 9) we can see that the parametric and nonparametric fits are relatively, close, with the exception being that the nonparametric fit is significantly more peaked (this was generally the case).

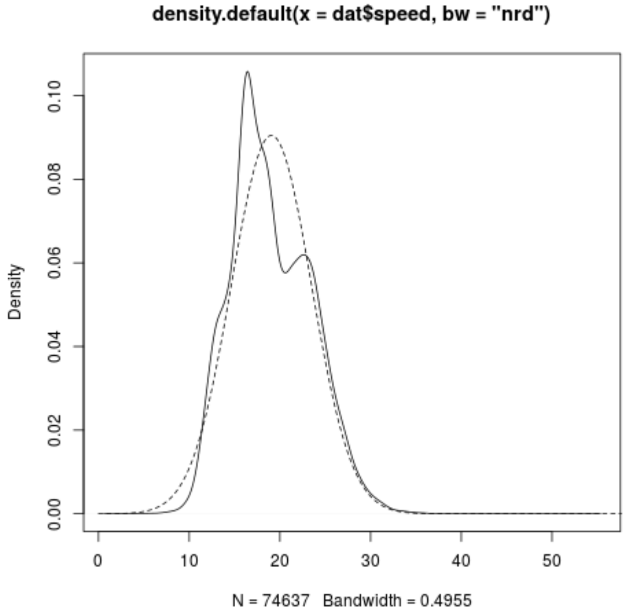


Figure 8: Region 13 parametric (dashed) and nonparametric (solid) densities

4.2 Regression Analysis

In an attempt to improve on the parametric models we wanted to fit a nonparametric regression to the data. To do this we chose to model the data using a multi-dimensional local linear regression. This model makes very weak assumptions about the underlying function (depending upon the desired convergence rate either that it is Lipschitz Continuous or that it has two bounded derivatives). Originally we considered using a generalized additive model (GAM) since local linear regressions begin to degrade in higher dimensions

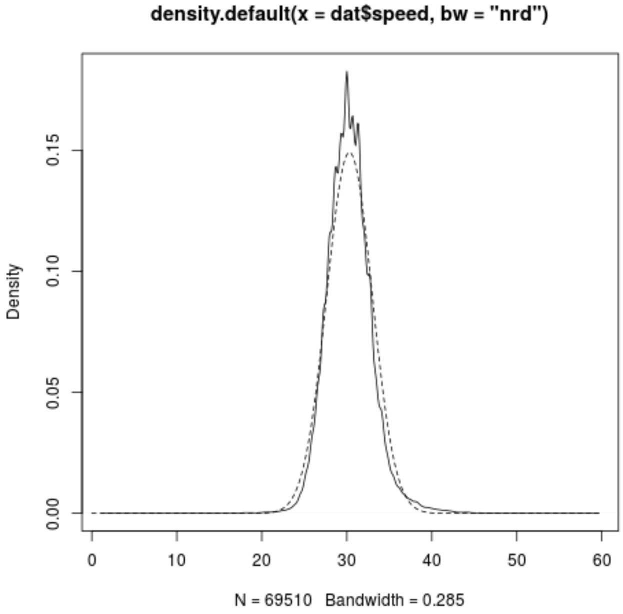


Figure 9: Region 22 parametric (dashed) and nonparametric (solid) densities

and we were planning on performing a higher dimensional regression. However, several of the models that we wanted to experiment with involved interaction terms which GAMs explicitly disallow. As such, we decided to stick with a local linear regression, despite the curse of dimensionality.

Our methodology for generating a local linear model was very similar to our methodology in the parametric case: we partitioned the data set, did model selection using AIC, and computed the training/test error. The one addition is that we implemented cross validation for selecting the optimal bandwidth. Unfortunately due, to computation and time constraints, we were unable to use cross validation during model selection; instead we used a heuristic value for the bandwidth (obtained through manual experimentation) and used the AIC value for this fit.[3] Once we had selected a model we performed cross validation to select the optimal bandwidth before training the final model. Table 1 contains the test and training error for the optimal fit for each region. In general, the nonparametric models seemed to outperform the parametric models; for example, if we look at region 12 (Figure 10) we can see that the nonparametric fit produces significantly lower test and training error than the parametric models. We discuss the comparison between these two more below.

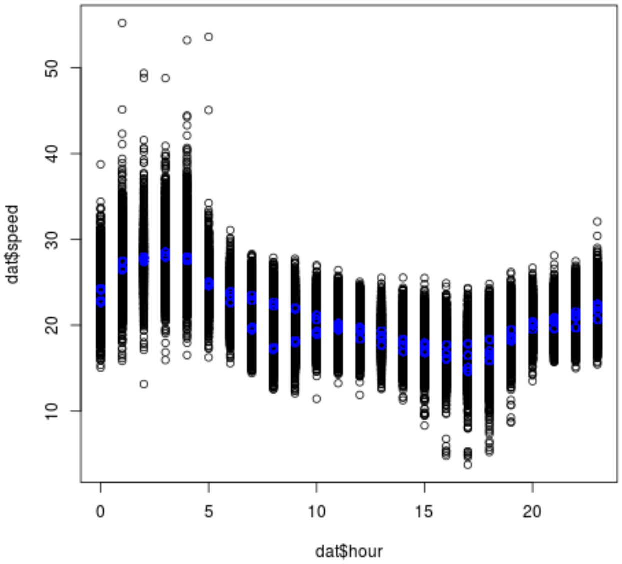


Figure 10: Region 12 nonparametric regression

5 Discussion

5.1 Density Estimation

It should be clear that the nonparametric density estimate is a much better tool for this data set than the parametric density estimate. In part this is because we were unsure of what distribution should be chosen when constructing a parametric density model; another part is that the data doesn't seem to follow any well known distribution and that the distribution seemed to vary from region to region. Although we could make further attempts to parametrize the density estimate (using the kernel density estimates to lead our parametrization), we believe that the

[3] We could have used leave one out cross validation instead of AIC during model selection, but we wanted to keep it consistent with our parametric analysis, and the two are asymptotically equivalent.

estimates produced by the KDEs are sufficient for our current goals.

5.2 Regression Analysis

Looking at the results from both parametric and non parametric analysis it seems that there were two types of regions: one group of regions was fit relatively well by the models in this paper (generally leading to a r^2 between 0.4 and 0.5 for a non linear model), and the other group on which all models failed miserably. Most of the regions fell into the first group, but regions 2, 16, 17, 22, 25, 27, 28, and 29 all fell in the second group. All of these regions, except 16, fall on the geographic periphery of the measurements. 16 is an interesting corner case because it seems to exhibit low hourly variation (see Figure 11), but still has very high variance; the other independent variables in our study show the same pattern. It seems that for this region, and possibly the other regions in group 2, our study is missing some key factors that influence traffic speed.

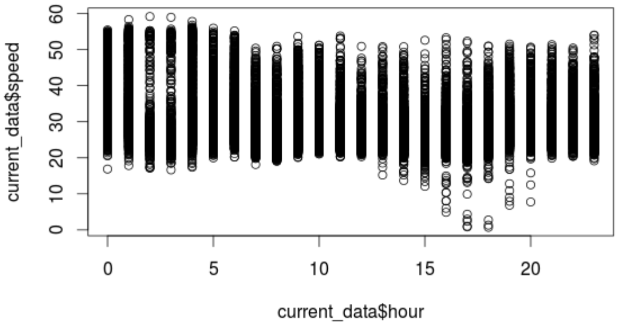


Figure 11: Region 16 speed vs hour

It is almost universally true that the cubic models outperformed the linear models; this is to be expected since cubic models are a superset of linear models and the data set is large enough that overfitting should not be a significant problem (at least with cubic models). The effectiveness of the nonparametric models is more complicated. In regions near downtown Chicago (12, 13, 14, 15) the nonparametric model significantly outperformed the cubic model on both the test and the train data. However, in other regions the nonparametric model performed at about the same level, or slightly worse than the cubic models. This could be indicative of the fact that traffic patterns are more complex in metropolitan areas.

5.3 Future Work

In this analysis we chose a relatively small subset of factors that we believed would impact traffic speed (time factors and sports games). However, it has become apparent that these factors are not sufficient to fully capture the variability of traffic speeds. The next step would be to expand the number of factors that are included in the analysis; possible factors could include weather (temperature, precipitation, etc.), gas prices, etc. Additionally the data set that we worked with contained only two years of data; a more expansive data set would allow us to better analyze long term trends and factors (population, unemployment, etc.). Finally, the analysis above was carried out ignoring the timeseries nature of the data. Another possible project would be to apply methods that take into account the fact that there is a time based dependence in the data.

ID	μ	σ^2 (plus, minus)	L Train	NL Train	NP Train	L Test	NL Test	NP Test	L r^2	NL r^2
1	24.70 ± 2.20E-02	8.41 (-2.80E-03,+2.97E-03)	6.13	3.89	3.6	6.2	3.91	3.65	0.28	0.54
2	29.02 ± 1.97E-02	7.05 (-2.30E-03,+2.43E-03)	6.75	6.08	5.92	6.88	6.13	6.08	0.05	0.14
3	24.97 ± 2.49E-02	11.77 (-3.75E-03,+3.97E-03)	8.6	4.66	4.81	8.47	4.58	4.75	0.28	0.61
4	23.06 ± 2.52E-02	11.55 (-3.76E-03,+3.98E-03)	7.77	4.22	4.01	7.84	4.27	4	0.33	0.64
5	25.90 ± 2.28E-02	9.16 (-3.03E-03,+3.21E-03)	6.92	3.99	3.7	7	4.01	3.77	0.24	0.56
6	24.29 ± 2.59E-02	12.48 (-4.03E-03,+4.26E-03)	9.18	5.49	5.32	9.27	5.62	5.48	0.26	0.56
7	23.83 ± 2.39E-02	10.36 (-3.38E-03,+3.58E-03)	7.6	4.23	4.13	7.93	4.42	4.38	0.27	0.59
8	22.66 ± 2.71E-02	14.19 (-4.49E-03,+4.75E-03)	9.77	5.95	5.13	9.79	5.96	5.15	0.31	0.58
9	25.76 ± 2.29E-02	8.95 (-3.01E-03,+3.19E-03)	7.28	4.38	3.99	7.38	4.52	4.15	0.19	0.51
10	24.51 ± 2.14E-02	8.86 (-2.80E-03,+2.96E-03)	6.73	4.28	3.62	6.86	4.36	3.67	0.25	0.52
11	24.29 ± 2.78E-02	14.99 (-4.74E-03,+5.01E-03)	11.17	5.79	5.67	11.14	5.66	5.64	0.26	0.62
12	20.78 ± 3.11E-02	18.21 (-5.84E-03,+6.18E-03)	12.13	6.48	5.05	12.21	6.53	5.06	0.33	0.64
13	19.07 ± 3.16E-02	19.44 (-6.14E-03,+6.49E-03)	13.89	5.92	3.32	13.73	5.84	3.25	0.29	0.7
14	26.58 ± 2.64E-02	12.95 (-4.18E-03,+4.42E-03)	10.4	6.2	5.04	10.31	6.09	4.94	0.2	0.53
15	27.51 ± 2.67E-02	13.60 (-4.32E-03,+4.57E-03)	10.67	7.42	6.17	10.81	7.57	6.35	0.22	0.45
16	29.11 ± 4.08E-02	28.75 (-9.60E-03,+1.02E-02)	23.52	21.64	20.65	23.09	21.44	20.58	0.19	0.25
17	31.36 ± 4.60E-02	34.83 (-1.19E-02,+1.26E-02)	32.18	23.32	20.63	32.58	23.72	21.24	0.07	0.33
18	27.05 ± 2.34E-02	10.52 (-3.33E-03,+3.52E-03)	8.03	4.36	4.22	7.74	4.09	4.05	0.25	0.59
19	26.34 ± 2.03E-02	7.97 (-2.52E-03,+2.66E-03)	6.11	3.95	3.85	6	3.84	3.77	0.24	0.51
20	27.55 ± 3.17E-02	19.15 (-6.10E-03,+6.46E-03)	14.86	12.81	12.32	15.37	13.38	12.99	0.22	0.33
21	26.54 ± 2.39E-02	10.40 (-3.39E-03,+3.59E-03)	8.06	5.53	5.22	8.07	5.53	5.26	0.23	0.47
22	30.31 ± 1.99E-02	7.16 (-2.34E-03,+2.48E-03)	6.59	5.53	5.21	6.83	5.73	5.45	0.08	0.23
23	24.41 ± 1.99E-02	7.66 (-2.42E-03,+2.56E-03)	5.38	3.65	3.31	5.5	3.77	3.46	0.3	0.53
24	26.04 ± 2.21E-02	8.95 (-2.90E-03,+3.07E-03)	6.78	4.8	4.45	6.79	4.78	4.46	0.24	0.46
25	31.40 ± 2.11E-02	7.20 (-2.49E-03,+2.64E-03)	6.85	5.78	5.67	7.16	5.96	5.96	0.05	0.2
26	30.33 ± 2.82E-02	15.09 (-4.82E-03,+5.10E-03)	12.17	9.16	7.85	12.24	9.09	7.79	0.2	0.4
27	31.06 ± 2.25E-02	9.18 (-3.00E-03,+3.18E-03)	8.58	8.26	7.76	8.92	8.6	8.2	0.06	0.1
28	34.94 ± 4.61E-02	26.55 (-1.04E-02,+1.11E-02)	26.04	24.82	31.48	26.48	25.1	31.71	0.02	0.06
29	23.97 ± 3.46E-02	19.75 (-6.75E-03,+7.17E-03)	16.67	13.77	11.45	17.08	14.17	11.87	0.16	0.3

Table 1: All numeric data captured in this paper. Includes sample mean and variance for each region as well as r^2, training, and test error for Linear (L), Nonlinear (NL), and Nonparametric (NP) models

6 Biblography

References

[1] "Chicago Traffic Tracker - Historical Congestion Estimates by Region — City of Chicago — Data Portal." Chicago. https://data.cityofchicago.org/Transportation/Chicago-Traffic-Tracker-Historical-Congestion-Esti/emtn-qqdi

[2] "LandOfBasketball.com, Information About the NBA Universe: Players, Teams and Championships." http://www.landofbasketball.com/

[3] "The Official Site of the Chicago Cubs." http://chicago.cubs.mlb.com/schedule/sortable.jsp?c_id=chc&year=2011

[4] "Downloadable CSV and XML Files." Sunshine Forecast Downloadable Data Files. http://www.repole.com/sun4cast/data.html#dataprior

[5] "Baseball Almanac.""Baseball Almanac. http://www.baseball-almanac.com/teamstats/schedule.php?y=2013&t=CHA

[6] "Hockey-Reference.com." Hockey-Reference.com. http://www.hockey-reference.com/teams/CHI/2011_games.html