USING PROBE DATA ANALYTICS FOR CHARACTERIZING SPEED REDUCTIONS AS WELL AS PREDICTING SPEEDS DURING RAIN EVENTS

by

William Logan Downing

A Thesis

Submitted to the Faculty of Purdue University

In Partial Fulfillment of the Requirements for the degree of

Master of Science

Department of Earth, Atmospheric, and Planetary Sciences

West Lafayette, Indiana

August 2020

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Wen-wen Tung, Chair

Department of Earth, Atmospheric, & Planetary Sciences

Dr. Darcy Bullock

Department of Civil Engineering

Dr. William S. Cleveland

Department of Statistics

Dr. Daniel Dawson

Department of Earth, Atmospheric, & Planetary Sciences

Approved by:

Dr. Daniel J. Cziczo

Dedicated to my sweet bride, Audrey Rae.
Without her prayers, encouragement, and support, this document may have never been completed.

ACKNOWLEDGMENTS

The author would like to thank the following:

- Dr. Wen-wen Tung, for her guidance and counsel over the past two years. Thank you for your patience with me throughout the process.
- The Joint Transportation Research Program, for allowing me access to their traffic datasets and their expertise in the realm of traffic studies.
- Howell Li, for his technical guidance and assistance over the past two years.
- Dr. Darcy Bullock, for his enthusiasm and help in allowing me to better understand the traffic world as well as his insight in tuning figures.
- Cassandra McKee, for her collection of the Aries crash data.
- Jairaj Desai, for his providing the construction work zones for this analysis' time period.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This study emphasizes the extreme variability present in traffic speed studies and the need for high resolution traffic and weather data in order to understand the interaction between traffic speeds and weather. I analyzed the impact rainfall has on roadway traffic speeds along I-65 in Indiana for the month of June 2018 and attempted to leverage this information to model and predict traffic speeds. To develop a statistical distributional understanding of the difference between traffic speeds under rain and non-rain conditions, Quantile-Quantile plots were generated in addition to fitting both scenarios to a gamma distribution. To compare how traffic speeds react to various precipitation intensities, boxplots were generated for comparison. Then, a baseline speed was defined using the median traffic speed under non-rain scenarios and was used to calculate speed reductions from the baseline at varying precipitation intensities. Finally, an XGBoost model is developed to attempt traffic speed predictions.

There are five key findings indicated by this study. First, the non-rain traffic speeds above the 5^{th} percentile are typically faster than their rain speed counterparts at comparable quantile levels. Second, traffic speeds exhibit a high amount of variance at varying precipitation intensity levels. Third, the gamma distribution does not suit traffic speed distributions at all locations and times of day under rain or non-rain scenarios. This result is consistent with previous findings that suggest traffic speed interactions are highly variable and based on a variety of factors that are hard to account for. Fourth, weekday traffic speeds from 1600 to 2200 UTC are the most strongly impacted across all regions during rain events seeing speed reductions of up to 10 mph, this is consistent with previous findings. Finally, the XGBoost model did not perform adequately in the configuration used in this study. The poor performance of the XGBoost model was somewhat anticipated as this study did not have access to traffic volume information and instead leverages proxy variables to account for this. The findings of this study demonstrate the need for finer scale studies on traffic—weather interactions and provides methodology that can be extended to other weather and traffic datasets.

CHAPTER 1. INTRODUCTION

According to the Federal Highway Administration (FHWA), there are roughly 5.8 million crashes every year, 21% of which are weather-related. Of this 21%, 46% are directly attributable to rainfall (Federal Highway Administration, 2018). Rainfall impacts traffic conditions in a variety of ways: visibility can be greatly hindered as intensity increases, pavement friction can be reduced, and water can pool up if the intensity is high enough or road conditions are just right. These factors greatly impact the speed of traffic and results in an increase in risk for motorists. This study aims to improve our understanding of these relationships and hypothesizes that as precipitation intensity increases, there will be a corresponding trend towards lower traffic speeds. This study will also attempt to leverage this relationship for the sake of modeling and carry out a prediction of traffic speeds.

A 10-year average for the United States from 2007 to 2016 showed rainfall was directly attributable to 556,151 crashes, 212,647 injuries, and 2,473 deaths (Federal Highway Administration, 2018). While the cost of a crash is extremely variable, a rough idea of cost can be calculated using the comprehensive crash unit Property Damage Only (PDO) cost. The FHWA lists the comprehensive crash unit PDO cost as being roughly $7,400 in 2001 (Harmon, Bahar, & Gross, 2018). Adjusting for inflation using the Bureau of Labor Statistics inflation adjustment tool (U.S. Bureau of Labor Statistics, n.d.) shows the comprehensive crash unit PDO cost for a crash is roughly $10,900. This equates to over 6 billion US dollars yearly on average in PDO costs due to crashes attributable to rainfall. This calculation leaves out the cost of human lives lost to rainfall related crashes.

From both a quality of life and cost perspective, traffic and rainfall interactions deserve a thorough analysis in order to develop a deeper understanding of the interactions that occur to better allow mitigation of the impact of precipitation on motorists. High-resolution spatial and temporal precipitation intensity data is much more accessible than it has been in the past, thanks to the work of the National Severe Storms Laboratory (Zhang et al., 2016). High-resolution spatial and temporal traffic data has also come a long way thanks to crowdsourcing of telematics information from a variety of sources sold by companies like INRIX (Kim & Coifman, 2014). Combining high-resolution traffic and weather data may shed light on the interaction between weather and traffic in new ways that have yet to be uncovered in the literature. Many previous studies do not assess

the weather component at a high enough resolution. A deeper understanding of these interactions lends itself to a variety of applications. Autonomous vehicles and mapping systems could readily make use of more detailed knowledge of weather—traffic interactions to make more informed decisions for driving as well as route choices.

While the interaction between traffic speeds and precipitation intensity seems somewhat intuitive as many have experienced the phenomenon of speeds decreasing as precipitation begins and intensifies, the literature on this subject is not comprehensive. An example of a situation in which traffic speeds were reduced in the presence of intense precipitation can be seen in Figure 1.1 for mile marker 34.4 on Highway 465 in Indiana. The left figure shows typical traffic patterns on Tuesday, June 19th, 2018. In the left-most figure, a significant traffic slowdown is pictured around 1700 EST, indicating typical afternoon rush traffic speed reductions. On Tuesday, June 26th, a heavy rain event was recorded around 1100 EST. Traffic speeds were shown to significantly decrease during the rain event's time frame compared with the previous Tuesday's traffic speeds.

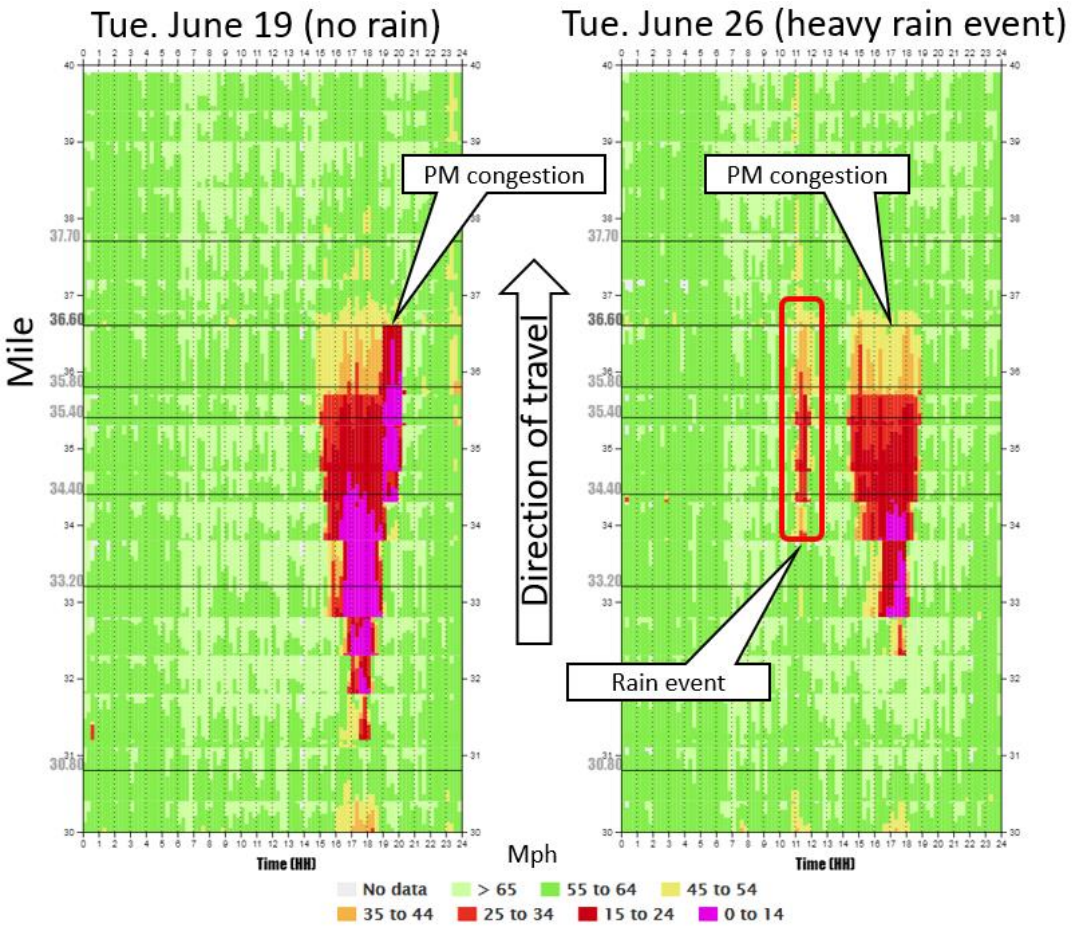


Figure 1.1. Traffic speeds near mile marker 34.4 on Highway 465 in Indiana. Typical PM congestion is contrasted with a rain event on June 26th, 2018.

The background information and literature review for the work carried out in this study will be covered in chapter 2, where previous research on weather's impact on traffic speeds will be discussed at length. Chapter 3 will discuss the data sources selected for this study. Chapter 4 will provide insight into the process selected for integrating traffic and weather data, engineering of features into the dataset, external data sources incorporated, the methods employed for analysis, and the modeling process. Chapter 5 will discuss the results from the analytics as well as the model performance. Chapter 5.1, in particular, will focus on the distributions of rain and non-rain traffic speeds, their likeness to a gamma distribution, data variability, and a measure of speed reductions under varying precipitation regimes. Chapter 6 will summarize the lessons learned from this study and provide some insight into how this work can be furthered in the future. The code developed for use in this study can be found at https://github.com/Foxhound013/MS-Thesis.

CHAPTER 2. BACKGROUND

Many studies have been performed over the years discussing the impacts of rainfall, or more broadly weather, on traffic speeds. Some of the earliest research carried out discussing the impact of weather on traffic was Tanner (1952), who noted a negative correlation between traffic speeds and rainfall. Tanner's work was somewhat limited in extent due to the time at which it was conducted. Further work was carried out by Ibrahim and Hall (1994) using high-resolution double loop detector traffic data for measuring volumes and speeds at 30-second intervals. Ibrahim and Hall's weather data was derived from a local airport of unspecified distance from the two locations used in their study. Their findings were much more specific than Tanner's and seem to be a touchstone study for future studies regarding weather's impact on traffic speeds. Ibrahim and Hall limited their analysis to 10 am to 4 pm in order to control for free-flow traffic conditions. Their findings indicate that light precipitation results in a speed reduction of roughly 1.2 mph and under heavy precipitation, a speed reduction of 3.1 to 6.2 mph. Their analysis, while thorough, did not adequately comment on the impact of traffic speeds relationship with precipitation at all times of the day. In addition to this, their analysis reduces the weather component to a much simpler picture than reality. A nearest-neighbor interpolation of the weather component does not adequately address the variations exhibited in real-world weather scenarios.

Further work carried out by Smith et al. (2003) sought to expand upon the work of the Highway Capacity Manual (HCM) as well as that of Ibrahim and Hall. They provided a summary of the work carried out by the HCM as well as Ibrahim and Hall. Smith et al. rightly noted the importance of regionality in their work. They mention that drivers from different areas are more or less comfortable with driving in inclement weather based on where they're from. This observation seems to be missed in much of the other work surrounding this topic. The traffic data used in their study was at the 15-minute scale, which seems reasonable, but their weather data was at an hourly scale and interpolated with a nearest-neighbor scheme from a single airport three miles away. The disparity in temporal and spatial resolution is somewhat of a cause for concern, as mentioned with Ibrahim and Hall. They noted that precipitation should be treated as a continual range rather than a binary variable. They found that capacity reduction is far more dramatic than that of speed. In order to discuss speed reductions, they propose an interesting method for producing a baseline speed for comparison. The baseline speed for comparison used in their study

is an average of traffic speeds greater than 40 mph. The average traffic speed above 40 mph was intended to mitigate the impact of congestion on the analysis. They did not note any significant difference in speed reduction from light to heavy rain intensities (.25-6.25 mm/hr and 6.25 mm/hr and above, respectively), contrary to the findings of Ibrahim and Hall.

The studies discussed up to this point have all made use of a nearest-neighbor interpolation scheme, which does not accurately reflect the large variability of weather conditions on small spatiotemporal scales, especially with respect to precipitation. This could be addressed by accessing a finer resolution weather dataset and interpolating the values to the road segment level as done in the research by Sathiaraj et al. (2018). Sathiaraj et al. analyzed visibility, wind speed, temperature, and precipitation. They made use of Inverse Distance Weighting (IDW) interpolation to adequately determine precipitation intensity at the 48 locations selected for their study. The weather data used in this study was somewhat coarse with an hourly temporal resolution and four weather stations for the 135 square miles of the area covered by this study. The extreme upper value of the precipitation analyzed in this study was above 7.62 mm/hr, which doesn't adequately describe precipitation rates in other locations. Their findings were primarily focused on traffic volumes, which seem to be more sensitive to inclement weather conditions than that of speed as their volume reductions are larger than the speed reductions found by Ibrahim and Hall. Despite Sathiaraj et al. focusing on volumes, their methodology is novel and groundbreaking in traffic—weather interactions as it provides the most accurate handling of weather data that this study has encountered in the literature.

All of the studies encountered by this research have been focused on relatively low "high intensity" precipitation. These studies fail to reference much higher precipitation regimes, which are somewhat common in other locations. Some locations can experience significantly higher precipitation intensities that may have a larger impact on traffic speeds than what has been seen in the literature.

Traffic modeling with weather information incorporated is not new and has been studied by Tsirigotis et al. (2012). Their work focused on the use of traditional time series forecasting methods like Auto Regressive Integrated Moving Average (ARIMA), ARIMA with Exogenous variables (ARIMAX), Vector Auto Regressive with Exogenous variables (VARX), and Bayesian VARX (BVARX). The weather data used by Tsirigotis et al. again seems to be a nearest-neighbor approach, but their model achieves good performance. Their findings indicated that truck

percentage, volumes, and rainfall as exogenous variables in the model improved predictive capability.

While the impact of precipitation intensity on traffic speeds is not new to the literature, very few studies have studied the data with adequate spatial and temporal resolution for both weather and traffic data across large geographic regions. This study aims to fill the gap in the literature by providing a more thorough analysis than what has been accomplished thus far. In addition to this, the literature suggests that volume, capacity, and the relative percentage of trucks on the road segment are highly important variables (Tsirigotis et al., 2012), this study will not be able to take those variables into account due to the nature of the traffic data used here. To account for this a modeling approach that works on weak learners, or variables that only have a slight correlation to with the true value, will be utilized to see if meaningful predictions can be made.

CHAPTER 3. DATA SOURCES

The Joint Transportation Research Program (JTRP) maintains a database of traffic speed data for the state of Indiana. The traffic data used in this study is derived from this database. Precipitation intensity data is derived from the Iowa Environmental Mesonet (IEM) Archived Data Resources ("Iowa Environmental Mesonet MRMS Grib Archive," n.d.), which maintains an active archive of a few MRMS variables.

3.1 INRIX Traffic Speeds

The traffic data used in this study is derived from the INRIX traffic speeds dataset. INRIX aggregates many sources of telematics data, commercial and personal, into representative segments of roadway with a temporal resolution of 1-minute. Other commercial datasets like INRIX exist from platforms like Here. In recent years Open Street Maps has begun to offer an open-source traffic data alternative. This study has not explored the viability of these other datasets for traffic weather interaction studies. The focus for this study was on liquid precipitation events along I-65 in Indiana from the southernmost point in Indiana near Louisville, KY to the northernmost point near Gary, IN, as can be seen in Figure 3.1. The segments consist of a start and endpoint with associated latitude and longitudes. It is sufficient for the purposes of this study to select the starting position of each segment and perform the analysis with that data point rather than working with the line segment. Only the northbound segments have been selected for analysis in this study.

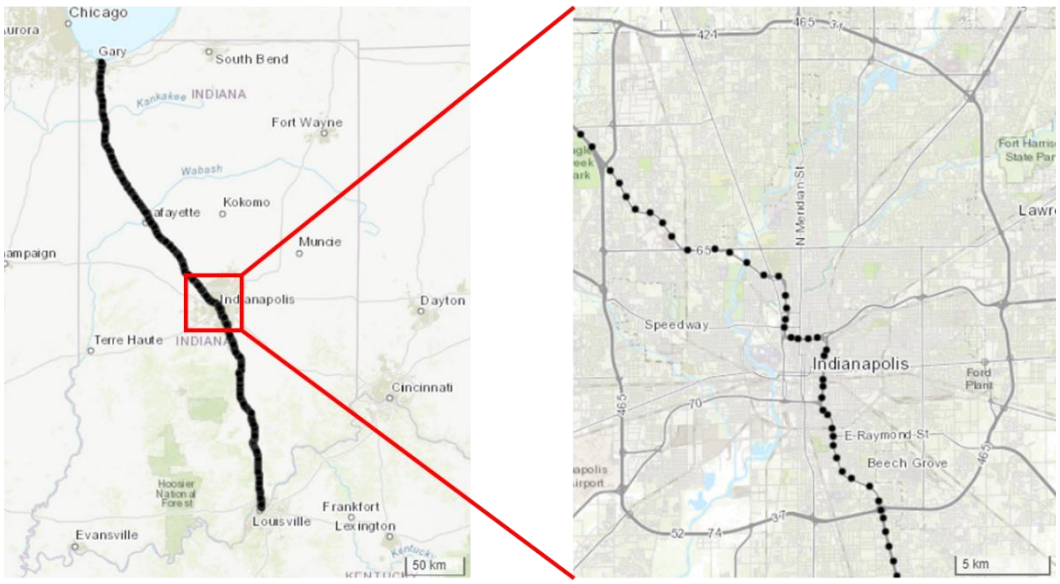


Figure 3.1. Northbound segments of I-65 in Indiana zoomed in to the Indianapolis region.

The data provider includes a confidence score at each time step, indicating the reliability of the data at that time step. A score of 30 constitutes high-confidence data and implies that the reported data is based on that segment's real-time data. A score of 20 constitutes medium-confidence data, indicating that the speed is based primarily on historical data, most likely the average speed. Finally, a confidence score of 10 constitutes low-confidence data, indicating that the reported speed is based solely on historical data, most likely the reference speed for that segment (Center for Advanced Transportation Technology, 2012). An analysis of the confidence scores available for June 2018 reveals that 96.77% of the data is high-confidence data, 3.23% is medium-confidence data, and 0% is low-confidence data. A review of the medium-confidence data indicates that the medium confidence data is not spatially clustered in any one location but occurs across all segments. Given the overwhelming percentage of high-confidence speed data, a negligible percentage of medium-confidence data, and a complete lack of low-confidence data, the dataset is complete. Given that the medium confidence data does not make up a large portion of the dataset and is not representative of reality, it will be discarded in this analysis. It should be noted that the INRIX data exhibits a weak temporal relationship in which there is a slight peak in medium confidence data around 6 UTC, or 2 am EST.

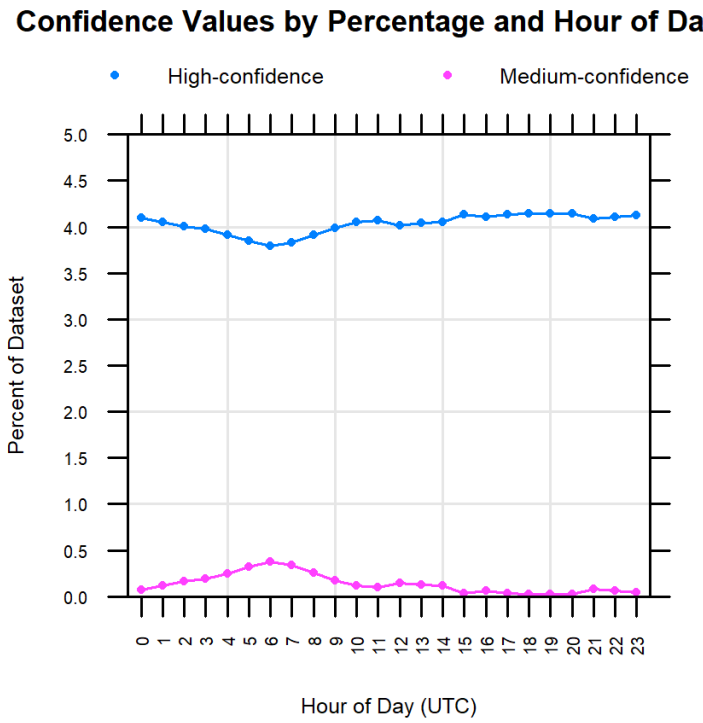


Figure 3.2. Percentage of confidence scores by hour of day.

3.2 MRMS Precipitation Intensity

The weather data used in this study is derived from the National Severe Storms Laboratory's (NSSL) Multi Radar Multi Sensor (MRMS) system. This system aggregates roughly 180 operational radars across the conterminous United States. It integrates them with atmospheric environmental data, satellite data, and lightning and rain gauge observations in order to develop a variety of weather products (Zhang et al., 2016). The variable of interest, for this research, in this dataset, is the precipitation rate. Zhang et al. calculates the precipitation rate using varying Rain Rate-Reflectivity (R-Z) relationships. Different R-Z relationships are implemented for stratiform rain, convective rain, and snow. This study is focused on June 2018 and therefore, will not include any precipitation rates calculated with the snow R-Z relationship. It should be mentioned that Zhang et al. note that the MRMS does not use polarimetric radar variables due to the need for further research at the time of publication.

The MRMS data was chosen for this study for a few reasons. The first reason for its choice was its high temporal and spatial resolution. The MRMS' precipitation rate is available at a 1 square kilometer spatial resolution and 2-minute temporal resolution. A high-resolution weather

dataset, both spatially and temporally, is a necessity for traffic—weather interaction studies. The second reason for its choice was the R-Z relationship calculations being varied based on the surrounding environment. This is critical as R-Z relationships are not constant and can vary significantly, as noted by Stout and Mueller (Stout & Mueller, 1968). The third and final reason for its choice is the vast amount of data that the system aggregates in addition to the ease with which it could be integrated into traffic management systems by a Department of Transportation (DOT).

CHAPTER 4. METHODOLOGY

This study will be carried out in two parts. The first portion of this study is intended to provide an understanding of the bulk properties of the relationship between traffic speeds and precipitation intensity. The second portion of this study will attempt to model the relationship between traffic speeds and precipitation intensity.

4.1 Speed Down-Sampling and Precipitation Interpolation

Two data sources are present in this research, the first being precipitation intensity and the second consisting of traffic speeds. Pre-processing of both datasets was required before analysis could begin. The traffic speed data was collected via JTRP's traffic speeds database and was down-sampled to a 2-minute temporal resolution in order to better align with the precipitation intensity. The down-sampling method used for the traffic speeds data was simply the speed at every 2-minutes. No averages are computed as the INRIX speeds data is an instantaneous measure of average traffic speed for that minute.

The MRMS precipitation data is a gridded dataset with a 1 square kilometer resolution for the continental United States. It is thus necessary to perform a clipping and interpolation step in order to subset the data to a manageable size and align it with the traffic speeds segment data. Clipping the precipitation data to Indiana was accomplished through the use of the wgrib2 tool (Climate Prediction Center, n.d.). A deterministic interpolation method, Inverse Distance Weighting (IDW) (Shepard, 1968), was leveraged to carry out the interpolation. A geostatistical interpolation method, Ordinary Kriging (ORK), was investigated for use but was ultimately discarded. It was, in theory, a comprehensive method, but in practice, it required too many assumptions and too much computational time. Ly and Charles found that on the daily scale, these two methods were comparable (Ly, Degré, & Charles, 2013).

The equation for IDW can be seen in Equation 4.1, and the accompanying weighting equation can be seen in Equation 4.2. In these equations, w_i refers to the weighting value assigned to each measurement in the interpolation field and $d(x, x_i)$ refers to the distance between the locations to be interpolated to, x, and the points to interpolate from, x_i. IDW requires two parameters, a power, p , to determine how strongly weighted surrounding measurements are, and

a max distance, D, of measurements to be considered in the interpolation. For the purposes of this study a power of 2 was chosen along with a max distance of 5 kilometers. It should be noted that while these chosen parameters may not be the most optimal choice but are adequate for the purposes of this study. Interpolation is carried out at each 2-minute time step for the month of June 2018. The results are joined with the traffic speeds data by segment ID for further analysis in this study.

Equation 4.1

$$\frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} \quad If\ d(x, x_i) > 0\ and\ d(x, x_i) \leq D$$

Equation 4.2

$$w_i(x) = \frac{1}{d(x, x_i)^p}$$

Unlike the traffic speeds data, there is missingness present in the precipitation intensity data. The missingness for the precipitation intensity data is particularly minor as it consists of two time slices (i.e., two observations). A simple naïve forecast from the previous time step is used to fill in the missing values in the precipitation intensity data.

4.2 Bulk Precipitation and Speed Analytics

A study of the bulk properties of the intersection between traffic speed and precipitation intensity is carried out to develop a deeper understanding of the overarching characteristics of traffic speeds in the presence and absence of precipitation for all segments of I-65 in Indiana. A majority of the dataset is dominated by typical conditions in which traffic speeds are relatively normal and there is no precipitation. A smaller subset of this data consists of traffic speeds that are slower than normal but are still in the absence of precipitation. Finally, an even smaller subset of this data consists of traffic speeds that are slower than normal and are in the presence of precipitation. The small sample size of data in which traffic speeds are slower than normal in the presence of precipitation presents a challenge in directly analyzing traffic speeds. The approach taken in this section of the study is an attempt at remedying this by reviewing the bulk properties of the data. This study recognizes that there are more prominent traffic dynamics variables such as traffic capacities, traffic volumes, human psychology, and other variables that likely play a more important role in determining the traffic speed.

Three tasks were carried out in this section, feature engineering, data augmentation, and analysis.

4.2.1 Feature Engineering

The sparsity of the information available in this study has given rise to the necessity of feature engineering. In order to obtain the clearest insights possible, two classes of features have been engineered into this dataset, spatial and temporal features. The first feature to be incorporated into this dataset is the spatial feature, region. The regions are defined somewhat arbitrarily to quantify the missing traffic capacity variable while time of day metrics were incorporated in order to derive the missing traffic volumes variable. The regions used in this dataset are as follows, Northern Indiana, Indianapolis, Louisville, and Rural. These regions, except for Rural, are represented in Figure 4.1. It should be noted that Rural is not outlined in Figure 4.1 as any region along I-65 in Indiana not defined by the red boundaries is considered as Rural in this study. The mile marker boundaries for the regions are given in Table 4-1. This study recognizes that these boundaries could be further discretized in order to improve homogeneity within regions. Using more generalized regions allows more precipitation occurrences to appear in each of the categories. Further division of the data could be detrimental to the sample size of traffic speeds in the presence of precipitation for some groups, given that this study is focused only on June 2018.

Table 4-1. Region mile marker boundaries

Region	Start Mile Marker	End Mile Marker
Northern Indiana	250.80	260.2
Indianapolis	105.73	130.56
Louisville	0.58	7.60
Rural	NA	NA

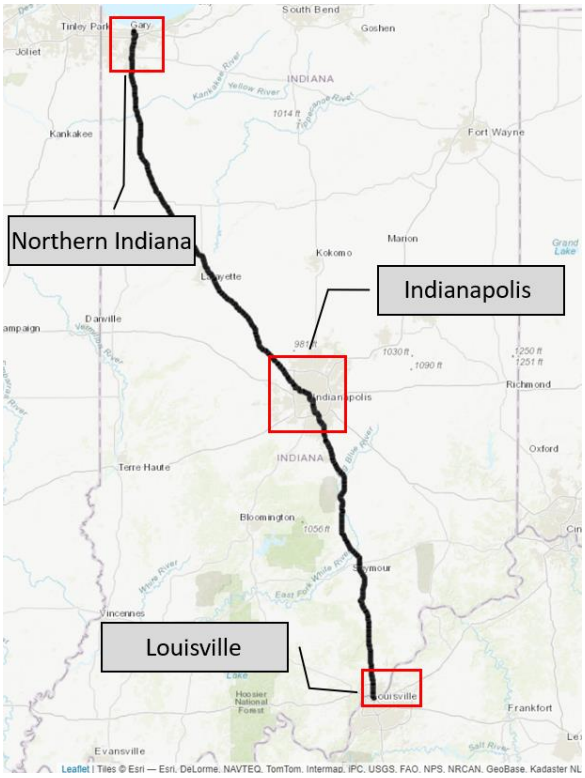


Figure 4.1. Region map of I-65

The temporal features incorporated into this dataset are as follows, hour range, day of week, and weekday/weekend. These features are based on the time stamp in the native dataset. These temporal divisions were selected to ensure the sample size in each division was adequate for analysis. A finer temporal resolution, such as the hourly scale, proved not to contain an adequate sample size at all time slices for analysis. It is for this reason that the hour of day and day of week were discarded in favor of an hour range and weekday/weekend feature.. The hour ranges selected for this study can be seen in Table 4-2.

Table 4-2. Hour ranges in UTC

Hour Range	Hours (UTC)
Morning	[4, 10)
Morning Rush	[10, 16)
Afternoon Rush	[16, 22)
Evening	[22, 4)

4.2.2 Data Augmentation

The dataset has been augmented with knowledge of crashes that occurred as well as with known construction zones in June 2018. In order to control for interference from crashes that occurred, crashes that occurred within DOT work zones have been filtered out of the dataset. Crashes were located by Cassandra McKee using the Automated Reporting Information Exchange System (ARIES) crash database in Indiana. Identified crashes have been filtered from the dataset by removing from the time at which a crash occurred to 30 minutes after and +/- 5 miles from the location in which it occurred. This is done to ensure that queuing issues and clean up associated with crashes are sufficiently removed from the dataset. Crashes, their times, and associated mile markers removed from the dataset can be seen in Table 4-3.

Table 4-3. Crashes removed from the dataset

Date	Times (UTC)	Mile Markers
06/08/2018	20:38 – 21:08	100 – 110
06/10/2018	19:30 – 20:00	56 – 66
06/12/2018	06:00 – 06:30	62 – 72
06/21/2018	17:30 – 18:00	172 – 182
06/21/2018	18:50 – 19:20	128 – 138
06/26/2018	14:50 – 15:20	110 – 120

Construction zones were marked as “Construction” or “Non-Construction” according to information provided by JTRP and INDOT. These construction/non-construction zones were determined by work zone reports compiled by Jairaj Desai at JTRP. The construction zones are defined in Table 4-4.

Table 4-4. Construction zone start and end mile markers

Construction Start Mile Marker	Construction End Mile Marker
8	16
50	68
141	165
167	176
197	207
229	253

4.2.3 Bulk Analytics

Once the feature engineering and data augmentation had been completed, the analytics were carried out as follows. A detailed analysis of the distribution of the rain and non-rain traffic speeds was carried out. A Quantile-Quantile (QQ) plot of the rain and non-rain speeds were produced in order to compare the traffic speeds at the same quantile level. The 5th, 25th, 50th, 75th, and 95th percentiles were marked in this plot in order to outline where a bulk of the deviations occur. Plots were conditioned on region, construction/non-construction, hour range, and weekday/weekend in order to bin sufficient sample sizes together but also separate the temporal and spatial components of these distributions. Further investigation of the distribution was made by comparing the rain and non-rain traffic speeds with a gamma distribution. The gamma distribution was selected as a candidate to test the fit of traffic speeds as the distribution is bounded at 0 and goes to positive infinity. The gamma distribution parameters were calculated according to Equation 4.4 and Equation 4.5 where μ and σ are the mean and standard deviation, respectively. To deal with extreme outliers in fitting the gamma distribution to the rain and non-rain traffic speeds, speeds less than 50 mph were removed from the data. The 50-mph threshold was chosen as most speed limits along I-65 are typically greater than 50 mph.

Equation 4.3

$$shape = \left(\frac{\mu}{\sigma}\right)^2$$

Equation 4.4

$$scale = \frac{\sigma^2}{\mu}$$

In order to understand how speeds reacted to different precipitation regimes, precipitation intensities were categorized into the following bins, [0, 0.01), [0.01, 2.5], [2.5, 5), [5, 10), [10, 20), [20, 30), [30, 40), [40, 50), [50, 60), [60, 70), [70, 80), [80, 150). Boxplots following the same conditioning scheme listed for the QQ plots are given. A sample size plot has been developed in addition to the boxplots in order to provide insight into how many samples are being represented by each box.

As a final method of understanding the bulk properties, the 50th percentile speeds were calculated by location against the precipitation intensity regimes used for the boxplots and conditioned on weekday/weekend, construction/non-construction, and hour range. The 50th

percentile speed under non-rain scenarios is also calculated using the same conditioning. The 50th percentile speed under varying precipitation regimes and non-rain scenarios is used to calculate a delta speed value. The delta calculation can be seen in Equation 4.5. In Equation 4.5, $Speed50r$ refers to the 50th percentile speed under the various precipitation regimes and $Speed50b$ refers to the 50th percentile baseline speed established by speeds in the absence of rain. Negative values indicate a slowdown from non-rain scenarios, and positive values indicate a speed up from non-rain scenarios.

Equation 4.5

$$Delta = Speed50r - Speed50b$$

4.3 Speed Prediction

The second step in this study was to attempt to develop a model relating traffic speeds to precipitation intensity so that, in theory, speed forecasts could be made in relation to precipitation intensity forecasts. A method for forecasting traffic speeds in addition to a baseline prediction to compare with would be required. The supervised learning method, eXtreme Gradient Boosting (XGBoost), was chosen as a suitable method for attempting to predict traffic speeds due to its reliance on weak learners as well as its extremely scalable nature (Chen & Guestrin, 2016). XGBoost is built to be quick and to leverage many CPU cores. This was ideal for this study as the dataset consisted of over 9 million observations. XGBoost is capable of classification as well as regression. This study will focus on its use as a tree-based regression technique.

Crashes and scores of 20 have not been removed for the purposes of modeling, as removing them would render the time series incomplete. The variables used in the model are as follows: traffic speed, a 16-minute lagged traffic speed, precipitation intensity, region, bearing, construction/non-construction, hour range, day/night, weekend/weekday, and day of the week. The traffic speed is the response variable, while the rest serve as predictors. The 16-minute lagged traffic speed has been added as an autoregressive component in order to help improve the model. A 10-minute and 60-minute moving average will be used as the baseline for which the XGBoost model will be compared and judged. Prior to modeling, it was required to carry out one-hot encoding on the categorical variables, such as region, hour range, day/night, day of week, weekend/weekday, construction/non-construction, and bearing. One-hot encoding of categorical

variables consists of creating a sparse matrix of 1s and 0s for each category. For example, one-hot encoding the day of week categorical variable would add 7 columns to the matrix. XGBoost has been designed with one-hot encoding in mind.

Given that the dataset is a time series, it was important to split the dataset so that the time series were still properly connected. The data was split into a training and testing set with the training set consisting of the observations before June 22 at 0 UTC (the first 70%) and the testing set consisting of the observations from and after June 22 at 0 UTC (the last 30%). Given the long running nature of the process, very few hyper parameters were tuned in this model. Two parameters in the XGBoost model were tuned; these include the boosting iterations and the eta parameter. A full list of parameter values used in this study can be seen in Table 4-5.

Table 4-5. XGBoost parameters used in study

Parameter	Values
Eta (Learning Rate)	0.1, 0.2, 0.3
N Rounds (Boosting Iterations)	200, 250, 300, . . ., 850, 950, 1000
Max Depth	6
Gamma	0
Columns Sampled by Tree	1
Minimum Child Weight	1
Subsample	1

The boosting iterations and learning rate were the only parameters tuned in this study due to time constraints. A 3-fold cross-validation was used to determine the best fitting model with the Root Mean Squared Error (RMSE) used as a measure of how well the model fit. Once the model is fit, Mean Absolute Error (MAE) will be the preferred method of this study for discussing model accuracy due to MAE being somewhat easier to interpret.

The learning rate is one method for preventing overfitting of the model from the training set by making each iteration, more conservative. The boosting iterations refer to the number of trees produced by the model. The max depth was left at its default value; it refers to how deep a tree can be grown or how complex a tree can become. The gamma value defaults to 0 and determines whether or not the tree can be split farther on a particular leaf node. The columns sampled by tree defaults to 1, which indicates that all columns should be included in the tree. The minimum child weight defaults to 1 and is used for determining whether to continue splitting a

leaf node. Finally, the subsample parameter defaults to 1 and is the ratio of the training set that is to be sampled for the current iterations tree building process.

A few segments will be selected for visualizing the time series of the predicted traffic speeds under the various methods in order to provide a visual view of how these models are performing and whether the added complexity and overhead of XGBoost is worth it in this scenarios. The XGBoost model will be compared with two baselines, a 10-minute moving average and a 60-minute moving average. The 10-minute moving average was chosen as a baseline as it is close to the 16-minute lagged speed that has been included in the model. The 60-minute moving average was included as a secondary baseline for comparison.

CHAPTER 5. RESULTS

5.1 Bulk Precipitation and Speed Analytics

A first step in attempting to understand the relationship between traffic speeds and precipitation intensity was a scatterplot. Given the size of the dataset in question, the results were unclear due to significant over-plotting. It should be noted that the occurrence of rain traffic speed scenarios is somewhat rare by comparison with non-rain traffic speed scenarios. The dataset for this study has roughly 8.65% of the data as rainy. Rather than plotting points individually, a distribution approach was taken in order to understand the data from a holistic standpoint. QQ plots are plotted by region, construction/non-construction, hour range, and weekday/weekend. The QQ plots can be seen in Figure 5.1 through Figure 5.6. It should be noted that the plots for construction zones in Indianapolis as well as Louisville have been omitted as there was no construction in these regions for June 2018. The 5^{th}, 25^{th}, 50^{th}, 75^{th}, and 95^{th} percentiles according to the non-rain distribution have been included as red solid and dashed lines in these figures in order to provide a sense of where in the distribution the deviations are occurring. A 45-degree line has been included in each panel of the figures to show precisely where the deviations in the two speed regimes are occurring. Points above the 45-degree line indicate that speeds are faster under non-rain scenarios at that quantile, while points below the 45-degree line indicate that speeds are faster under rain scenarios at that quantile.

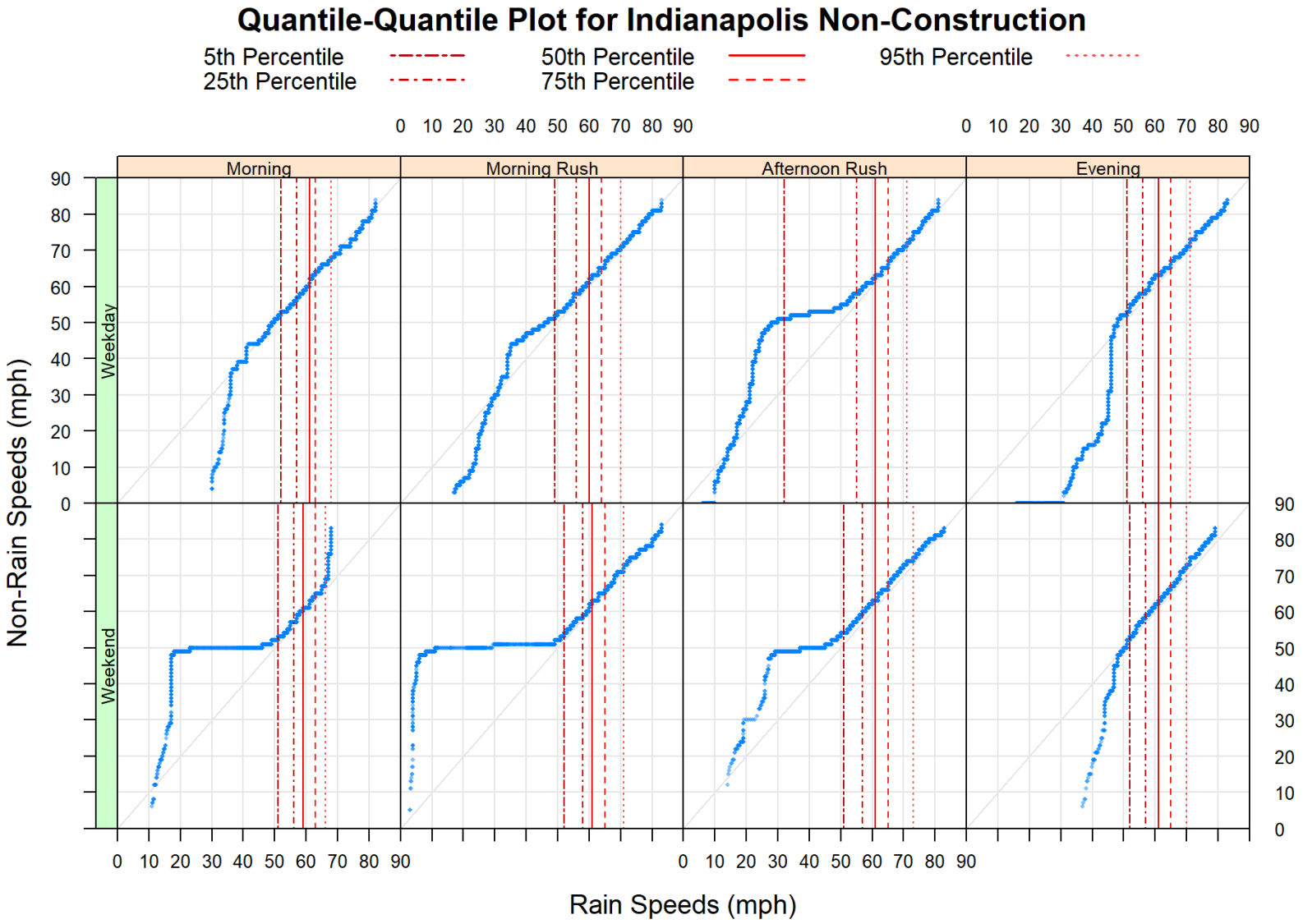


Figure 5.1. QQ plots of Indianapolis for non-construction, hour ranges, and weekday/weekend

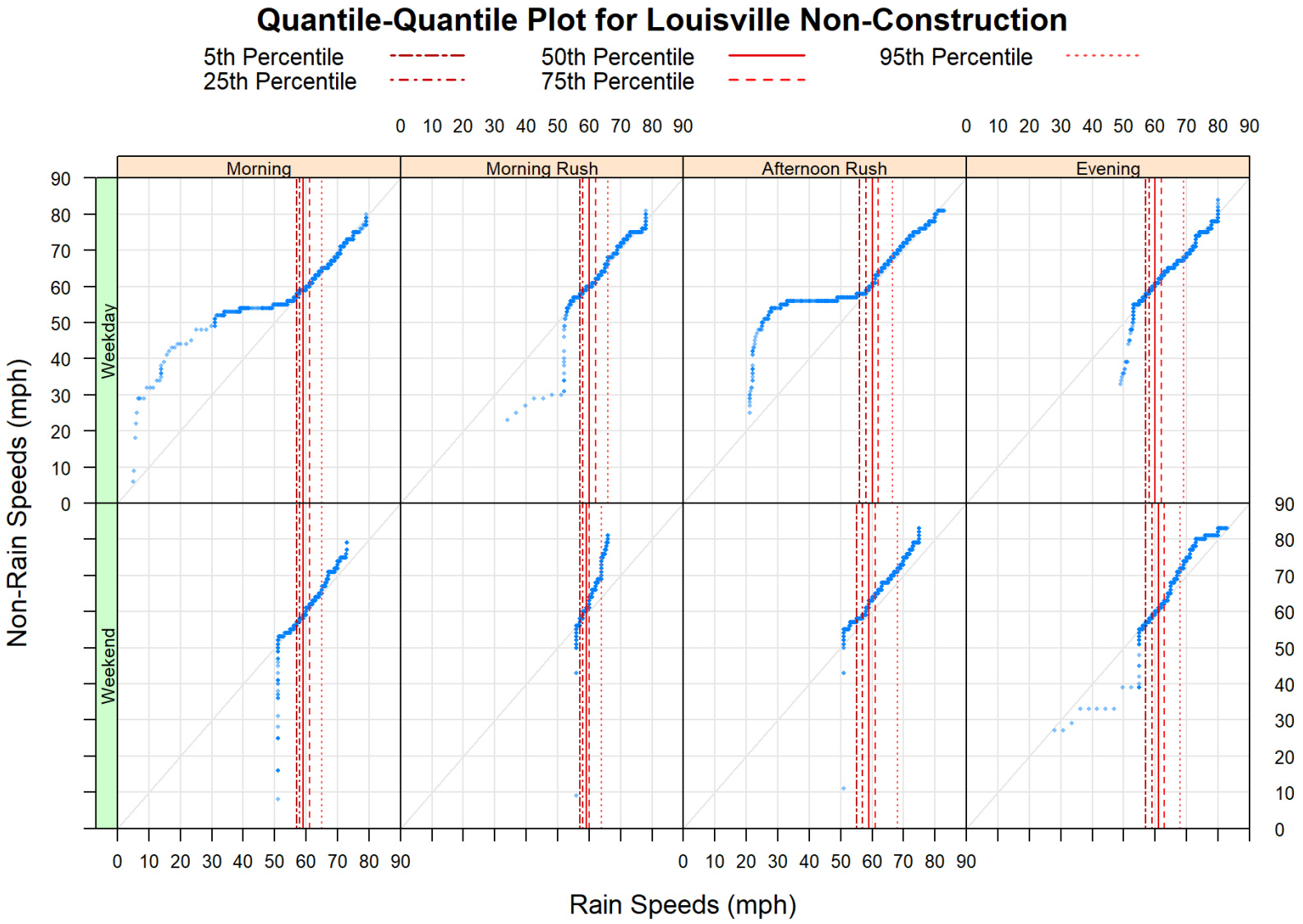


Figure 5.2. QQ plots of Louisville for non-construction, hour ranges, and weekday/weekend

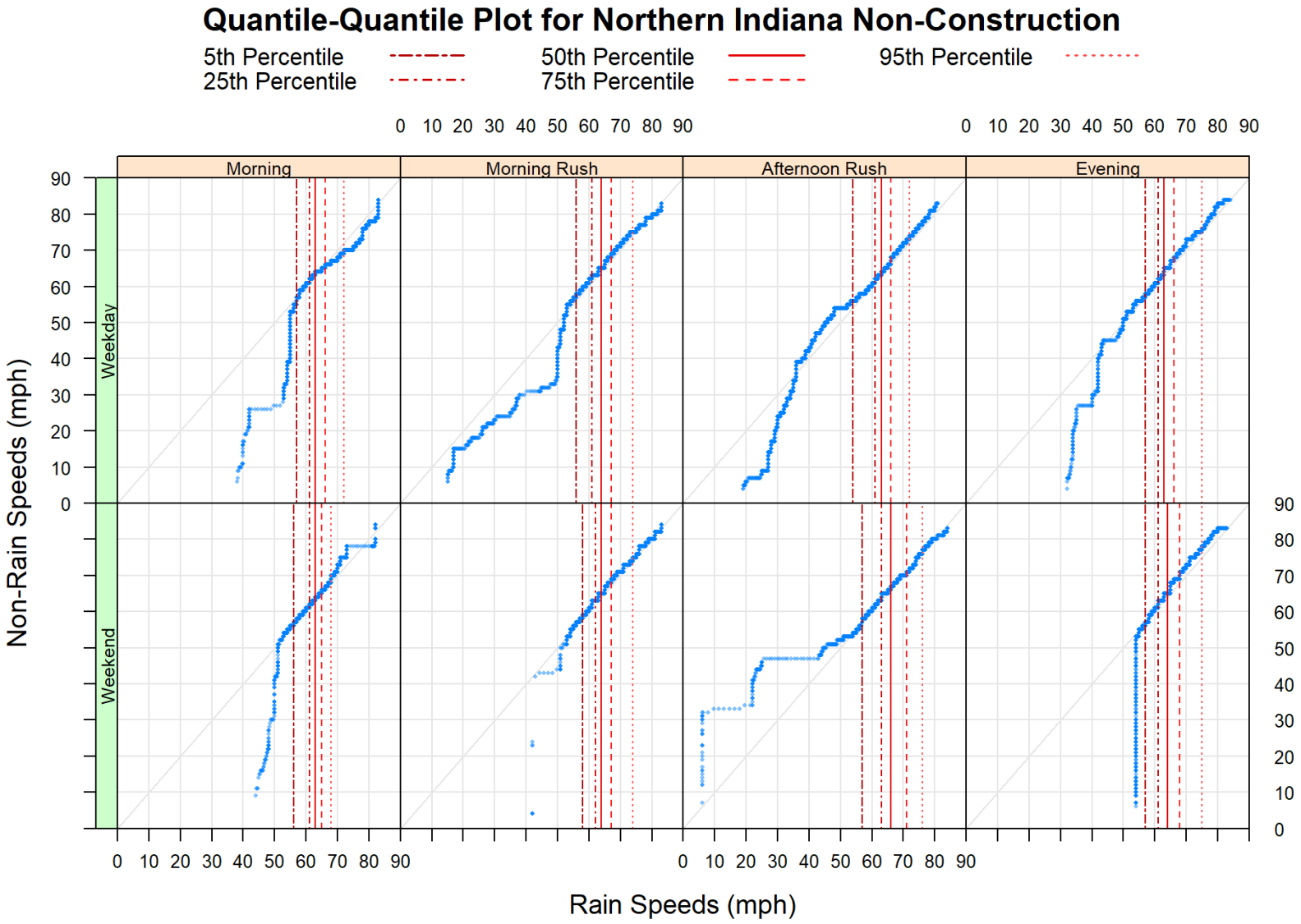


Figure 5.3. QQ plots of Northern Indiana for non-construction, hour ranges, and weekday/weekend

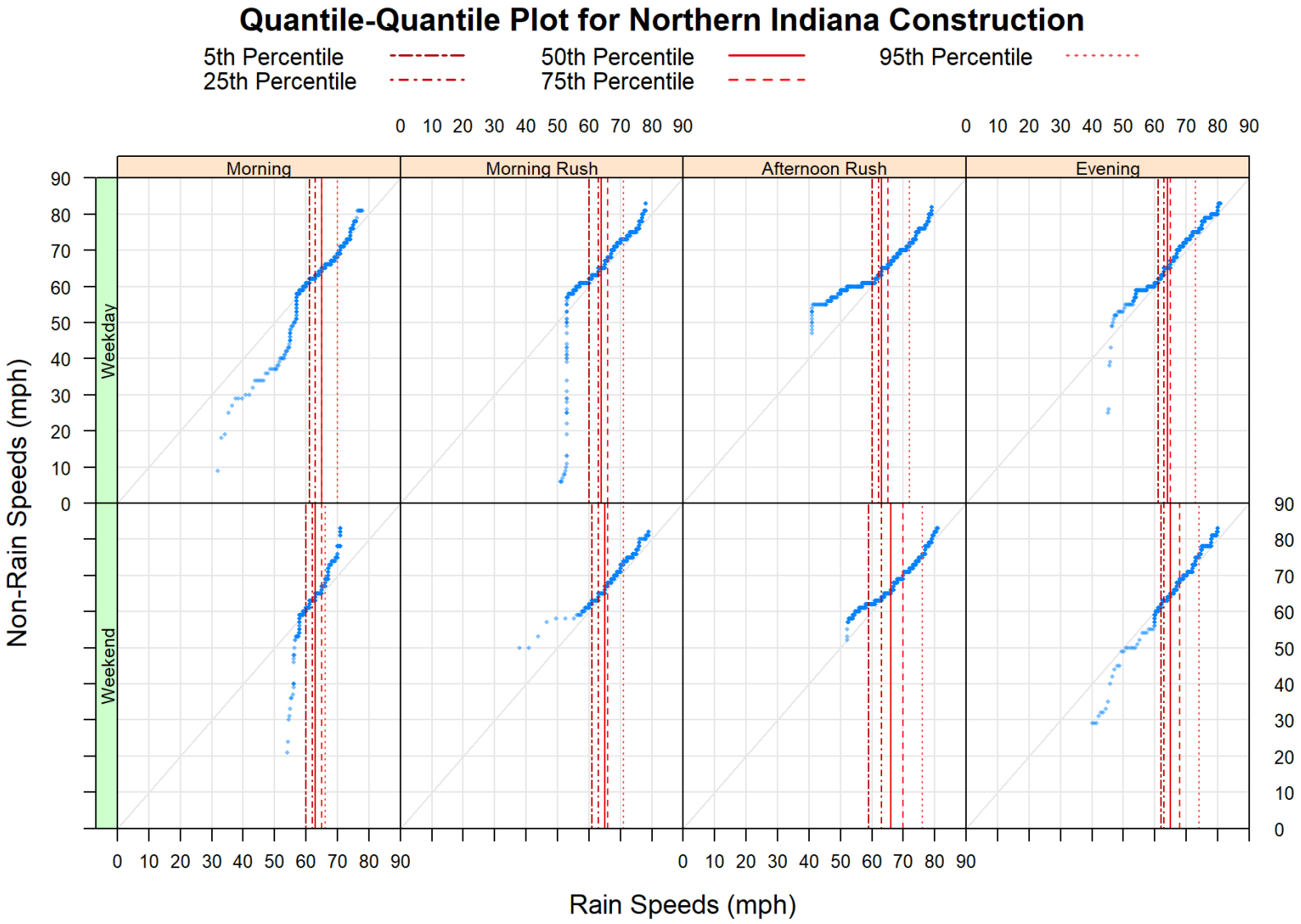


Figure 5.4. QQ plots of Northern Indiana for construction, hour ranges, and weekday/weekend

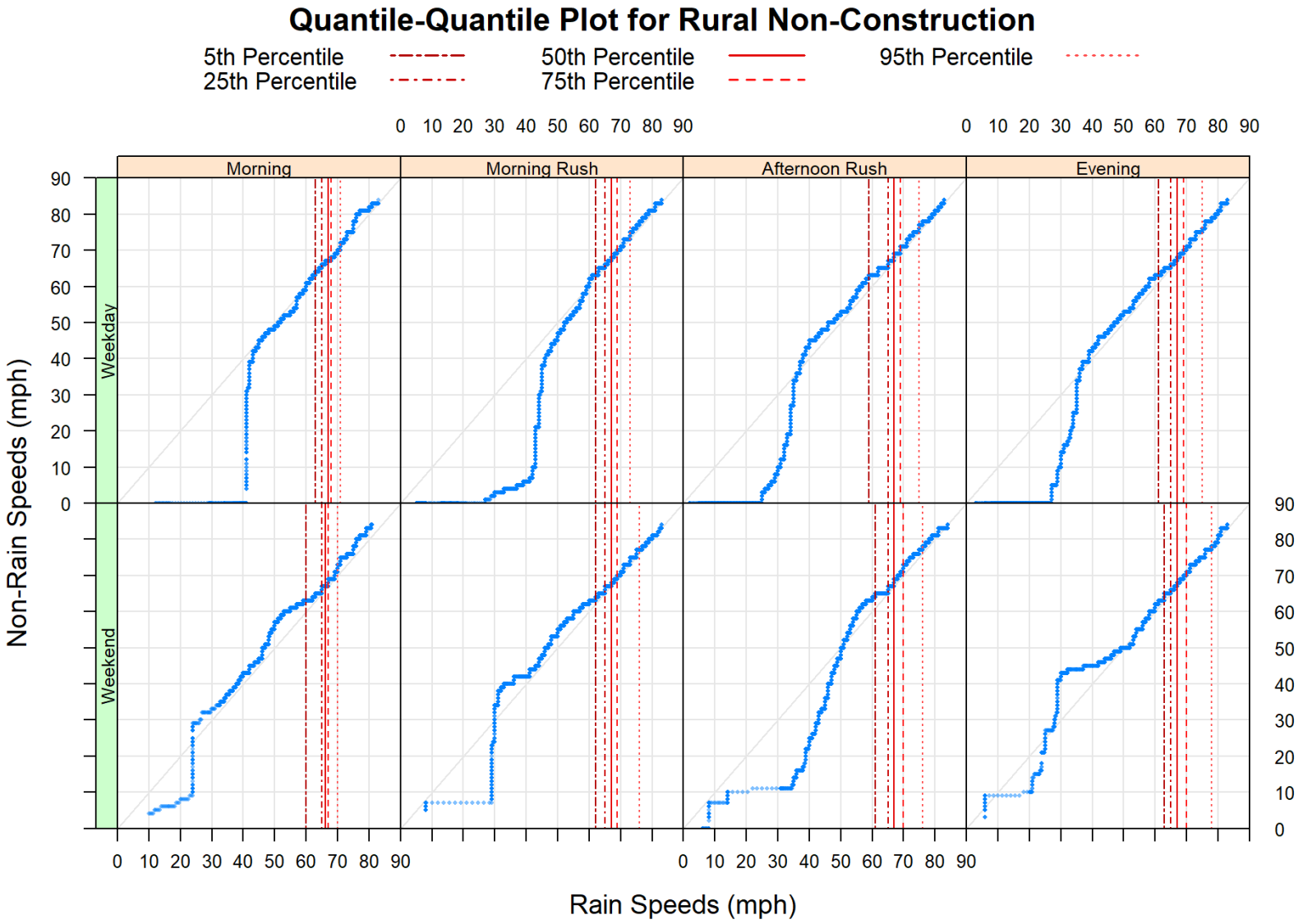


Figure 5.5. QQ plots of Rural areas for non-construction, hour ranges, and weekday/weekend

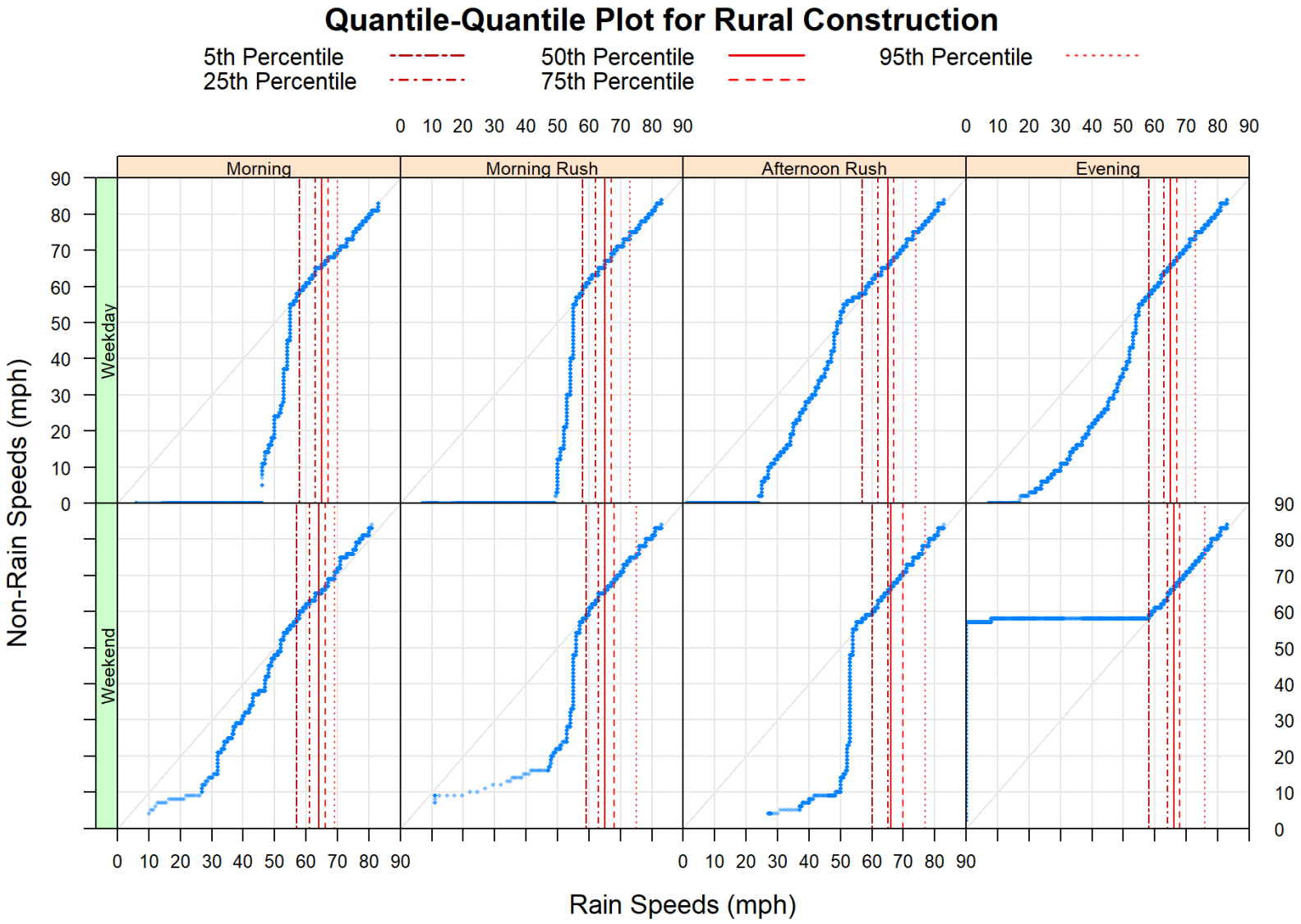


Figure 5.6. QQ plots of Rural areas for construction, hour ranges, and weekday/weekend

The results of the QQ plots are not as straightforward as one might hope. The reality of the situation is particularly nuanced as one considers each of the QQ plots in Figure 5.1 through Figure 5.6. Reviewing the situation in Indianapolis, Figure 5.1, above the 5th percentile, speeds are roughly equivalent with a slight tilt towards being faster under non-rain scenarios. The major deviations, as with most of the plots, occur at or below the 5th percentile. In the morning, morning rush, and afternoon rush time frames, the QQ plots show a hump that occurs during both weekdays and weekends towards non-rain speeds being a fair bit faster than rain speeds (with the exception of weekday mornings). In the case of weekday mornings and evenings as well as weekend evenings, the rain speeds are shown to be faster at the same quantile level. While this is somewhat counter intuitive at first, it is helpful to note that this is a very small proportion of the population.

Like Indianapolis, there are no construction zones to show in the Louisville region, as such only the non-construction zone is shown in Figure 5.2. The situation is a little different from that of Indianapolis. During the weekdays, the speeds above the 5th percentile mostly follow the 45-degree line. Below the 5th percentile, the morning and afternoon rush show speeds being faster under non-rain scenarios while the morning rush and evening time frames show speeds faster during rain scenarios. Speeds are indicated as being faster during rain scenarios on the weekend below the 5th percentile as well. Still, the speeds above the 75th percentile on the weekend indicate non-rain speeds as being markedly faster. It should be noted that the traffic dynamics of the Louisville region are substantially different than that of Indianapolis due to the makeup of the surrounding area. This may be one reason that the weekend time frame is so different from that of Indianapolis.

The Northern Indiana region for non-construction and construction zones can be seen in Figure 5.3 and Figure 5.4, respectively. For non-construction zones in Northern Indiana, the non-rain speeds seem to be faster below the 5th percentile in all situations, with the afternoon rush on the weekend being the one exception. Above the 5th percentile, speeds seem to be relatively comparable or a little faster under non-rain scenarios by and large, with the weekday morning being the one exception. This pattern, however, does not remain true for the construction zones shown in Figure 5.4. In the Northern Indiana construction zones during the morning weekday and weekend as well as the evening weekend, rain speeds are faster below the 5th percentile. Above the 5th percentile speeds are nominally the same. Interestingly the afternoon rush on the weekday and weekend as well as the weekend morning rush display significantly faster speeds below the

5th percentile under non-rain conditions. The construction zone speeds seem to lean towards faster speeds under non-rain conditions than non-construction zones do in Northern Indiana.

Finally, the Rural region QQ plots can be seen in Figure 5.5 and Figure 5.6 for non-construction and construction zones, respectively. The non-construction zone shows speeds to be roughly equivalent or slightly faster under non-rain scenarios until much lower in the distribution, typically well below the 5th percentile. The afternoon rush on the weekday is particularly notable as the non-rain speeds remain a fair bit faster than rainy speeds well past the 5th percentile. The same can be said of the weekday evenings and all time slots on the weekend aside from the afternoon rush. The construction scenario is not quite the same here, with rain speeds being faster below the 5th percentile. Above the 5th percentile, speeds are nominally the same. One oddity is the weekend evening speeds seem to be much faster under the non-rain scenario.

It should be noted as with all of these QQ plots, the sample sizes are important. With the weekend evening oddity in Figure 5.6, it is likely that this is an artifact of sample size issues. It should also be noted that despite how strange it might seem that rain speeds are faster in some cases at the same quantile level, this is typically below or well below the 5th percentile level. This indicates that a very minor proportion of the data is showing faster rain scenario speeds, whereas a majority of the data is showing comparable speeds or slightly faster speeds under non-rain scenarios. The QQ plots seem to support a weak relationship with traffic speeds and rain or non-rain scenarios.

In order to further investigate the distribution of the traffic speed data, comparisons with a Gamma Distribution have been made. Initial gamma distribution estimations can be seen in Appendix A. It was found that the initial gamma distribution estimations were greatly skewed by outliers. In order to refine the estimation of the gamma distribution, a threshold of 50 mph was used to see if the gamma distribution better fit the data under more typical speed regimes. The results of the gamma distribution fit can be seen in Figure 5.7 through Figure 5.12. As suggested by Dey et al., 2006, it has been shown that traffic speeds follow a normal distribution under homogeneous scenarios but deviate under heterogeneous conditions; comparisons to the normal distribution have been made in Appendix A to complete this portion of the analysis.

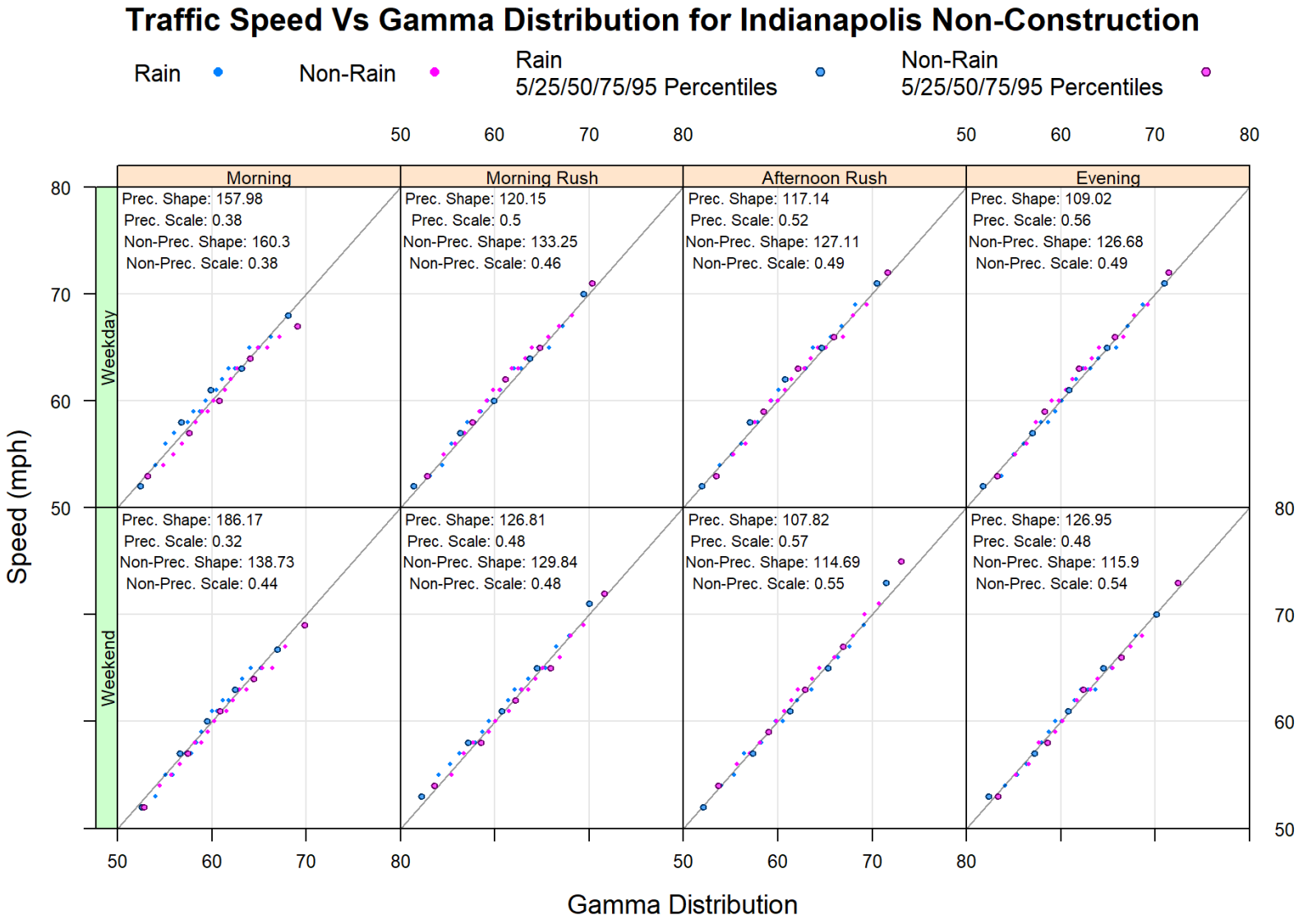


Figure 5.7. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Indianapolis, non-construction

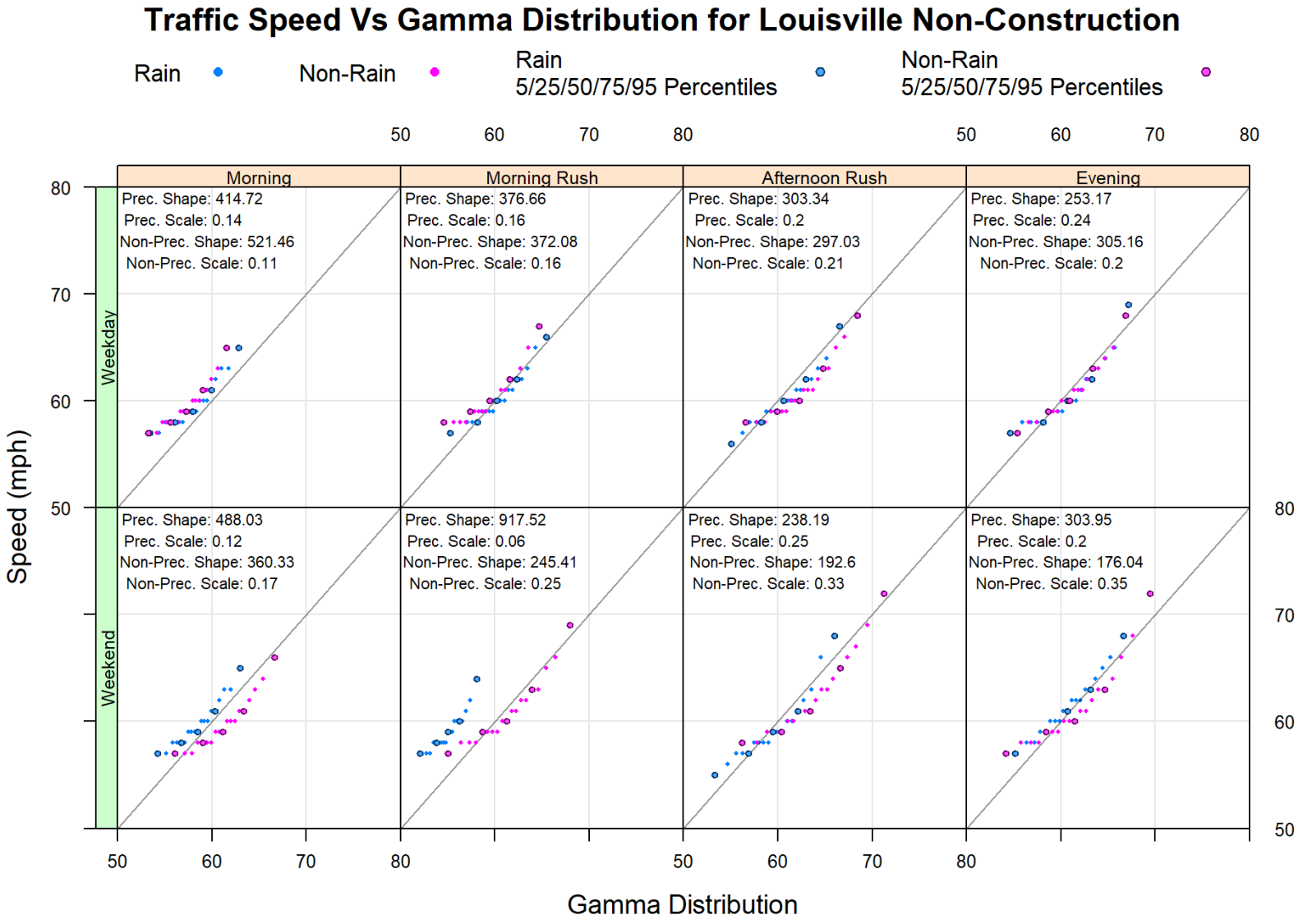


Figure 5.8. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Louisville, non-construction

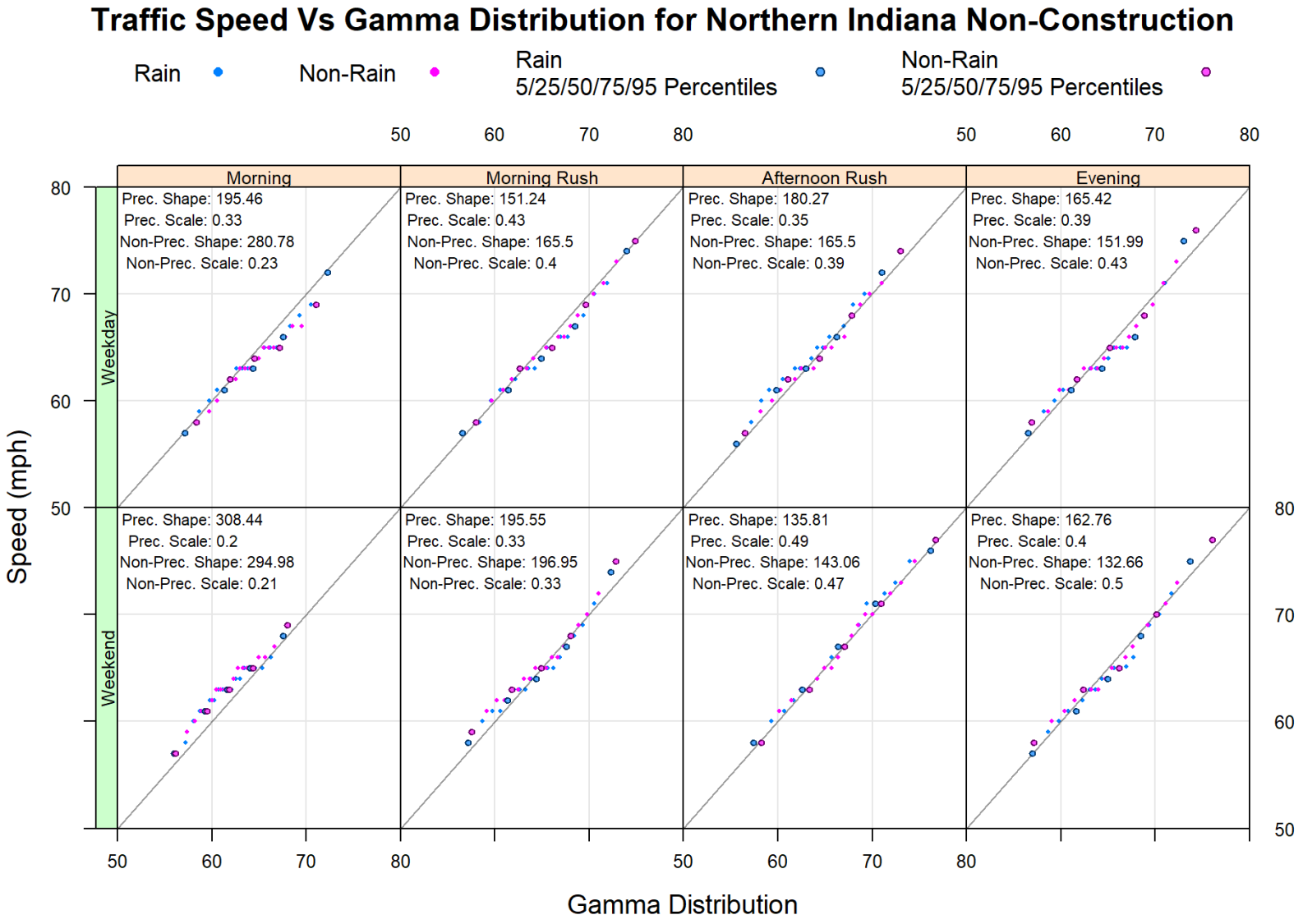


Figure 5.9. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Northern Indiana, non-construction

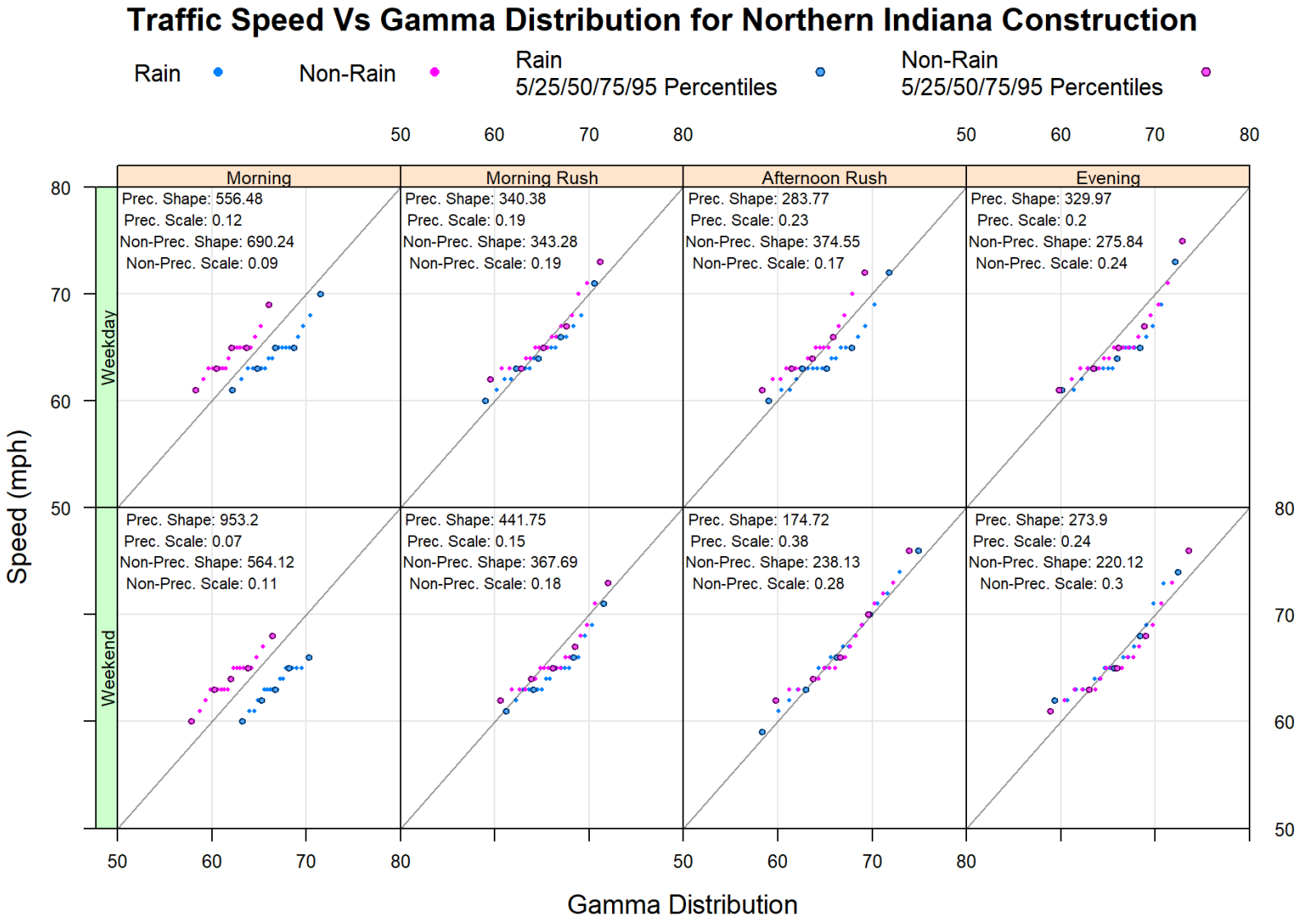


Figure 5.10. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Northern Indiana, construction

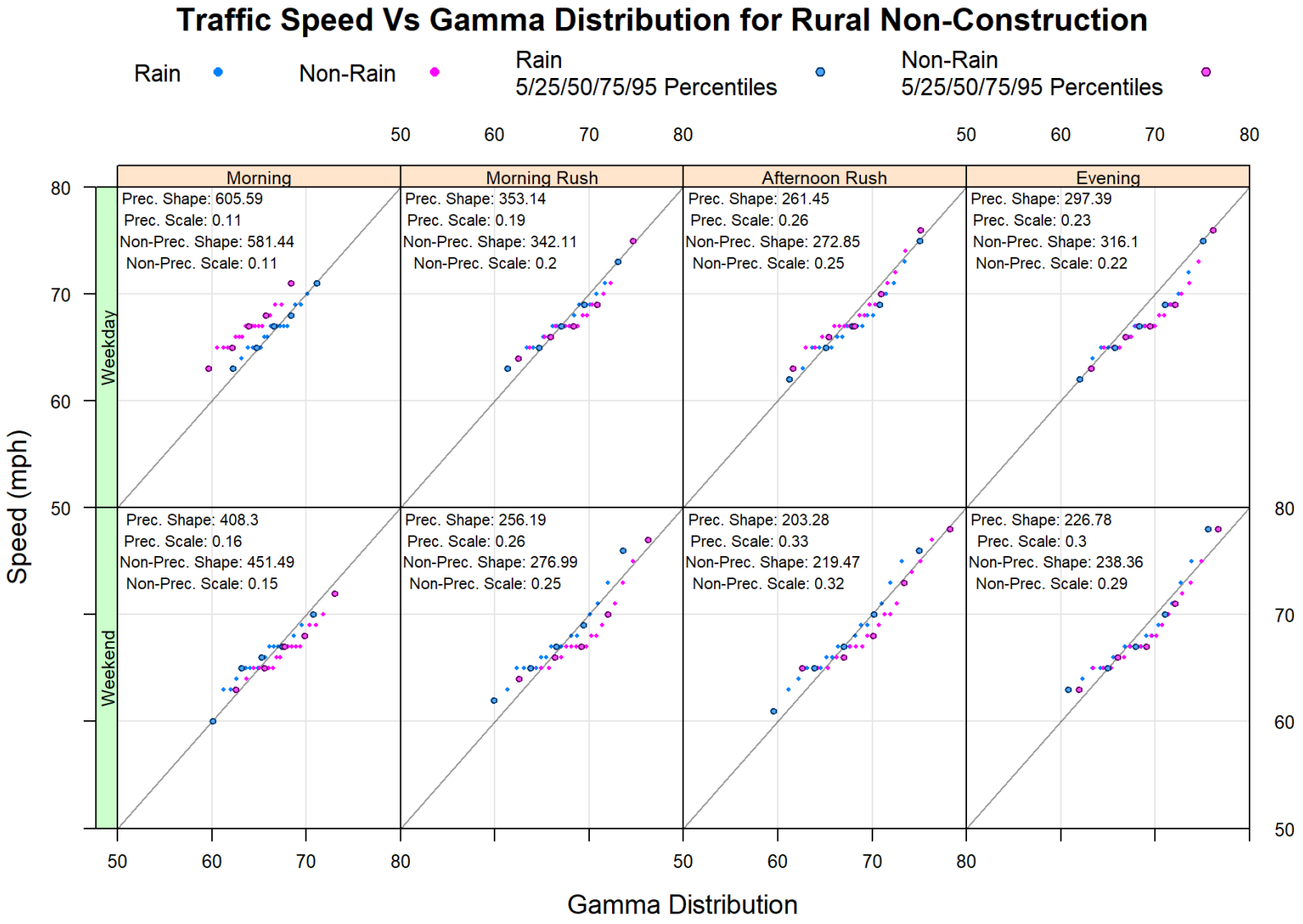


Figure 5.11. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Rural areas, non-construction

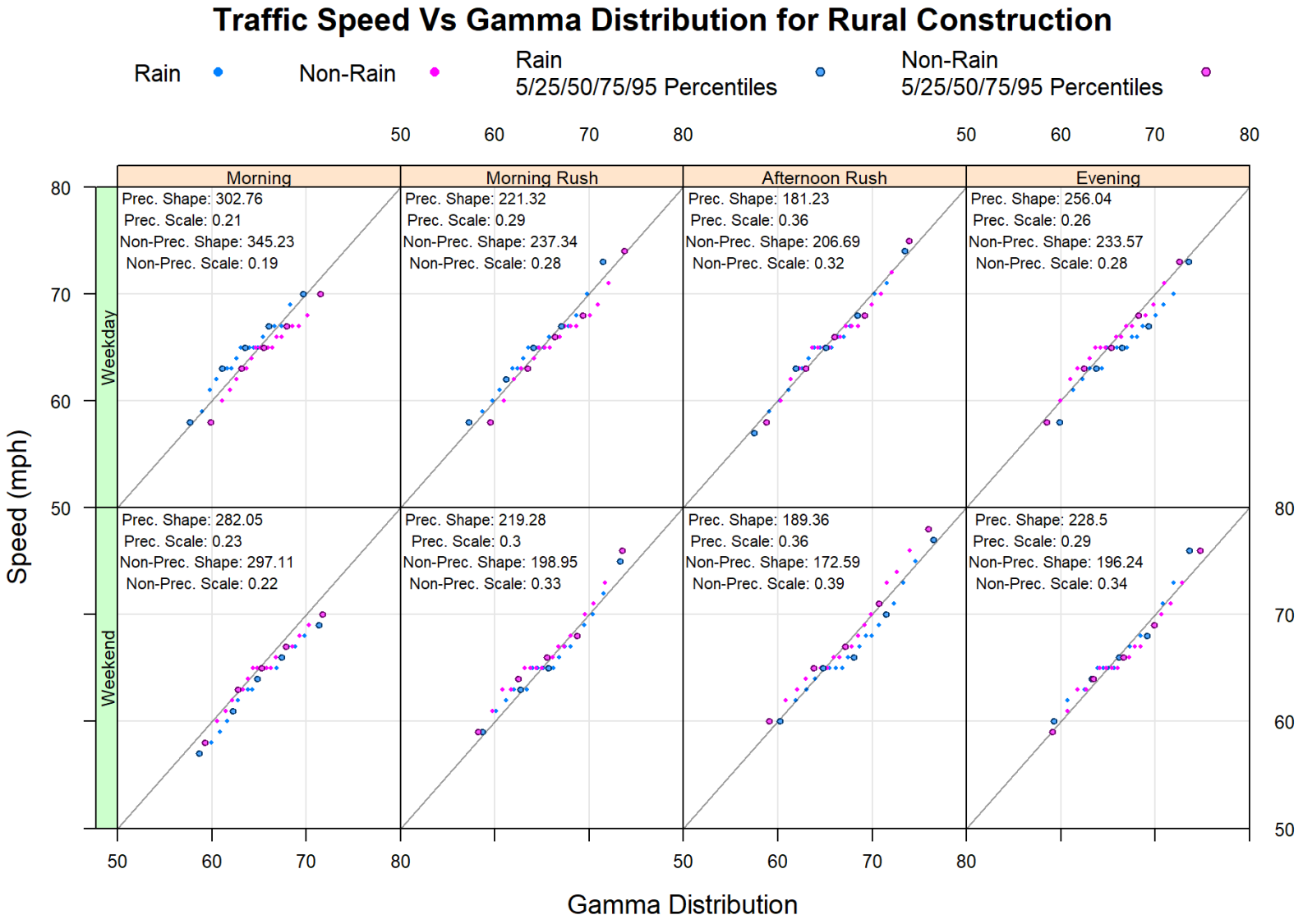


Figure 5.12. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Rural areas, construction

The results of the gamma distribution fit plots in Figure 5.7 through Figure 5.12 are somewhat mixed. In the Indianapolis, non-construction region shown in Figure 5.7, the gamma distribution seems to be overall a relatively good fit, and the separation between rain and non-rain gamma distribution is overall negligible. This was promising as the Indianapolis region appeared to be possible to model with a gamma distribution of the specified shape and scale parameters seen in Figure 5.7.

The Louisville non-construction region, as seen in Figure 5.8, was not nearly as promising with the gamma distribution not fitting well at any time slice, particularly on the weekends. There was much more of a separation between the separately fitted gamma distributions of rain and non-rain scenarios than what was seen in Indianapolis. The strong deviations from the 45-degree line as well as the curved nature of the distributions point to the gamma distribution being a poor fit for the Louisville region. This is the opposite of what we saw in the Indianapolis region.

Moving to the Northern Indiana region, the results are a little more nuanced than in Indianapolis or Louisville. The non-construction region shown in Figure 5.9, overall seems to show a good fit to the gamma distribution. The shape and scale parameters are much more reasonable than what was seen in Louisville, and the rain and non-rain scenarios are much more similar, much like Indianapolis. This changes significantly, though, when moving to the construction region in Figure 5.10. For the construction region, the weekend and weekday morning fits show a large separation between the rain and non-rain scenarios. The weekday afternoon rush also shows a separation of rain and non-rain scenarios but not to the same extent as the morning time frame. The overall fit to the 45-degree line is generally worse than the non-construction counterpart in Northern Indiana.

Finally, for the Rural non-construction and construction zones seen in Figure 5.11 and Figure 5.12, the results are still mixed. The non-construction traffic speeds seen in Figure 5.11 tend to have little to no separation for most time slices, with the exception of the weekday morning time frame. Overall, the gamma fit appears to be okay under the non-construction scenario but not as good as was seen in Indianapolis. A review of the rural area in the presence of construction, as seen in Figure 5.12, yields a slightly better fit to the gamma distribution. As with the non-construction gamma distribution fit, the morning time frame seems to display the largest separation between rain and non-rain gamma fits.

It would seem that the gamma distribution is only conditionally appropriate. There seemed to be some consistency in that non-construction zones, with the exception of Louisville, were better fit by the gamma distribution. As was noted by Smith et al. (Smith et al., 2003), there is a need to study the impacts of weather at varying locations to see if the weather-specific impact is region-specific. These gamma distribution fit plots seem to indicate that there is certainly a regional component involved.

Up to this point, precipitation intensity has not played a role in the visualizations and analysis produced by this study. To remedy this, boxplots as well as sample size plots have been generated with precipitation intensities, grouped as indicated in section 4.2.3. It should be noted that not all boxplots or sample size plots list each rain regime; this is due to there being no samples from the missing precipitation regime. These boxplots and sample size plots can be seen in Figure 5.13 through Figure 5.24. It should be noted that for the sample size plots, in many cases, the [0, 0.01) and [0.01, 2.5) precipitation ranges have a very large number of samples. It is for this reason that they have been excluded in order to allow for more differentiation in the lower sampled classes. The conditional grouping is carried out the same as was done with the ECDF plots previously. The boxplot whiskers are given as the magnitude of the 1^{st} and 3^{rd} quartiles multiplied by 1.5. Any data points that fall outside of this range are presented as outliers and shown as blue dots. The 50^{th} percentile is marked as a single black dot in the center of each box.

Three general observations can be derived from a review of Figure 5.13 through Figure 5.24. First, in many scenarios there seems to be a weak trend towards lower speeds as precipitation intensity increases. This observation seems to hold true across most locations. This observation is somewhat limited in the fact that there is a relatively small sample size as precipitation intensity increases at all locations. Second, there are many outliers, especially in the lower precipitation regimes. These outliers do start to fade away as precipitation intensity increases, but so does sample sizes for that precipitation regime. Finally, there is a significant amount of variance in the traffic speed amongst all precipitation regimes. This is particularly notable at higher precipitation regimes, though the 50^{th} percentile speed seems to show a lessening trend as precipitation intensity increases the values overlap significantly with lower precipitation intensity regimes.

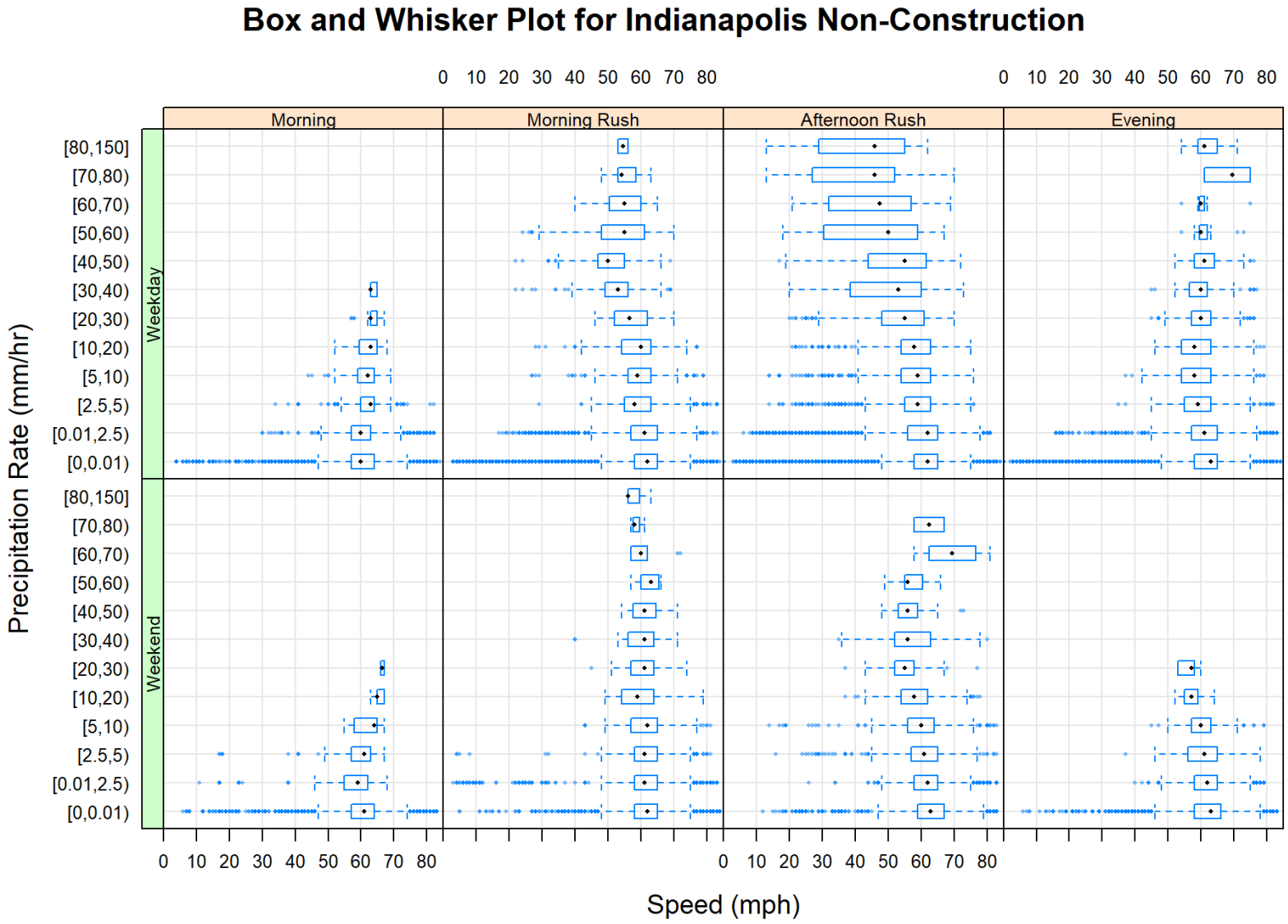


Figure 5.13. Boxplots of Indianapolis for non-construction, hour ranges, and weekday/weekend

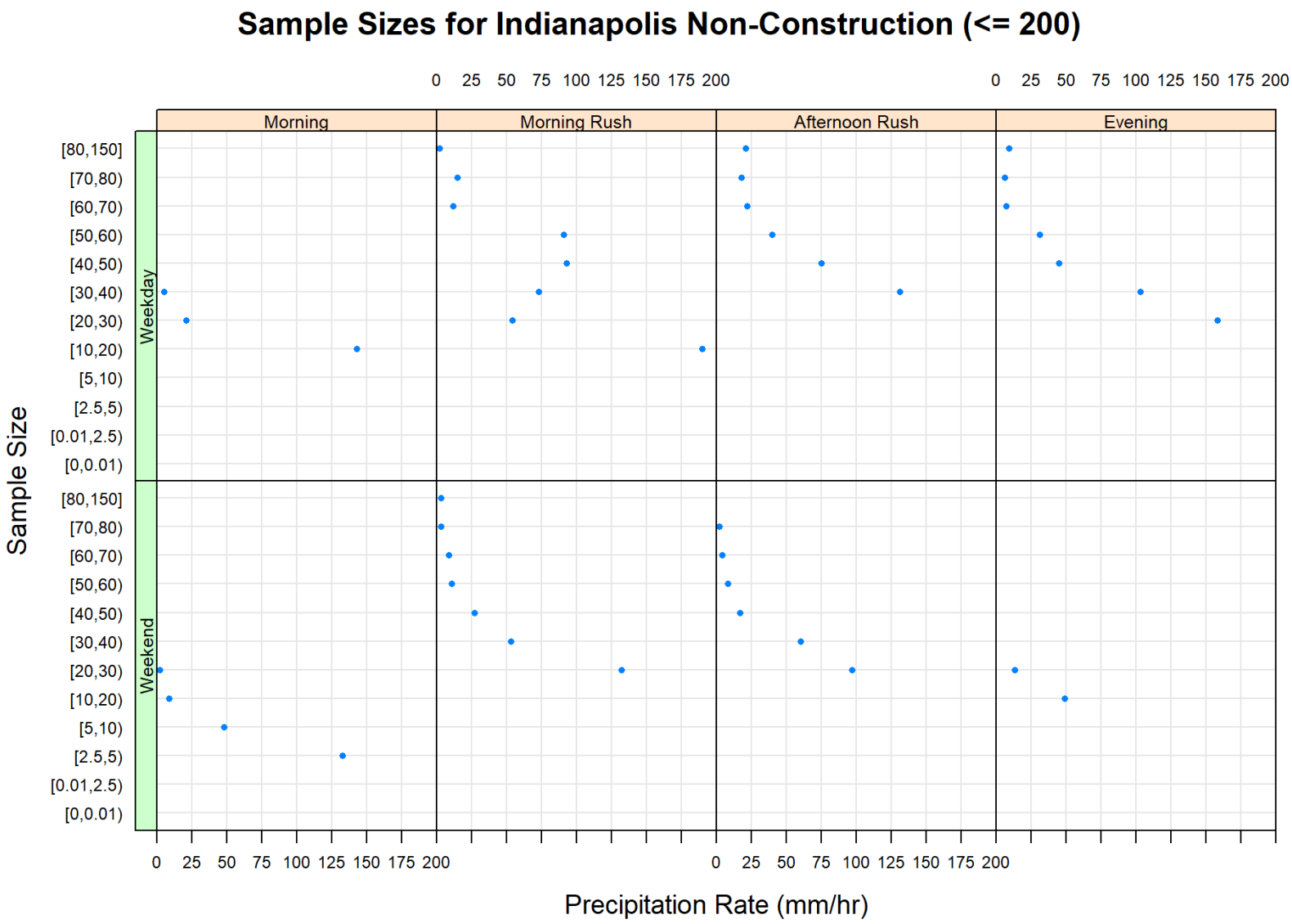


Figure 5.14. Sample size plots of Indianapolis for non-construction, hour ranges, and weekday/weekend

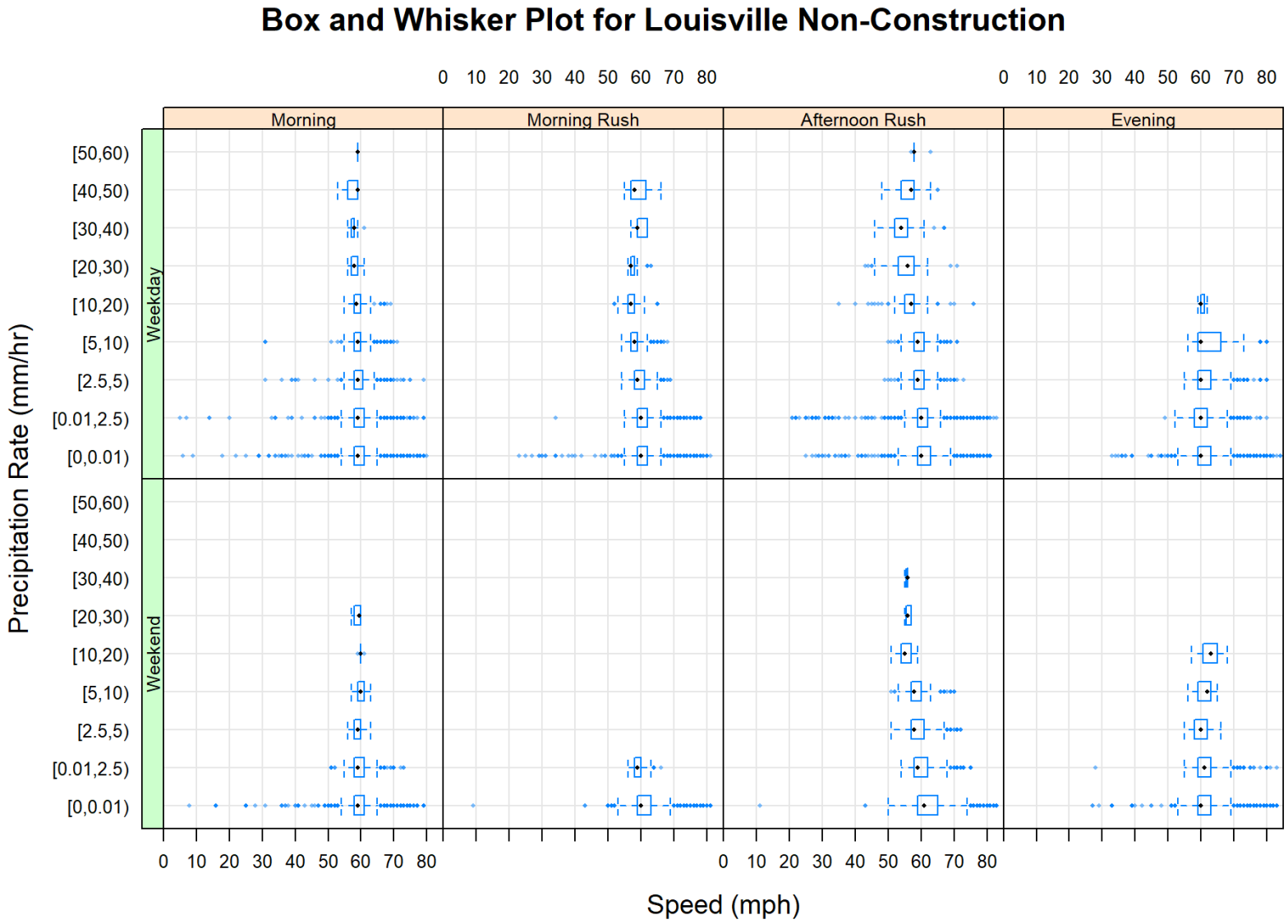


Figure 5.15. Boxplots of Louisville for non-construction, hour ranges, and weekday/weekend

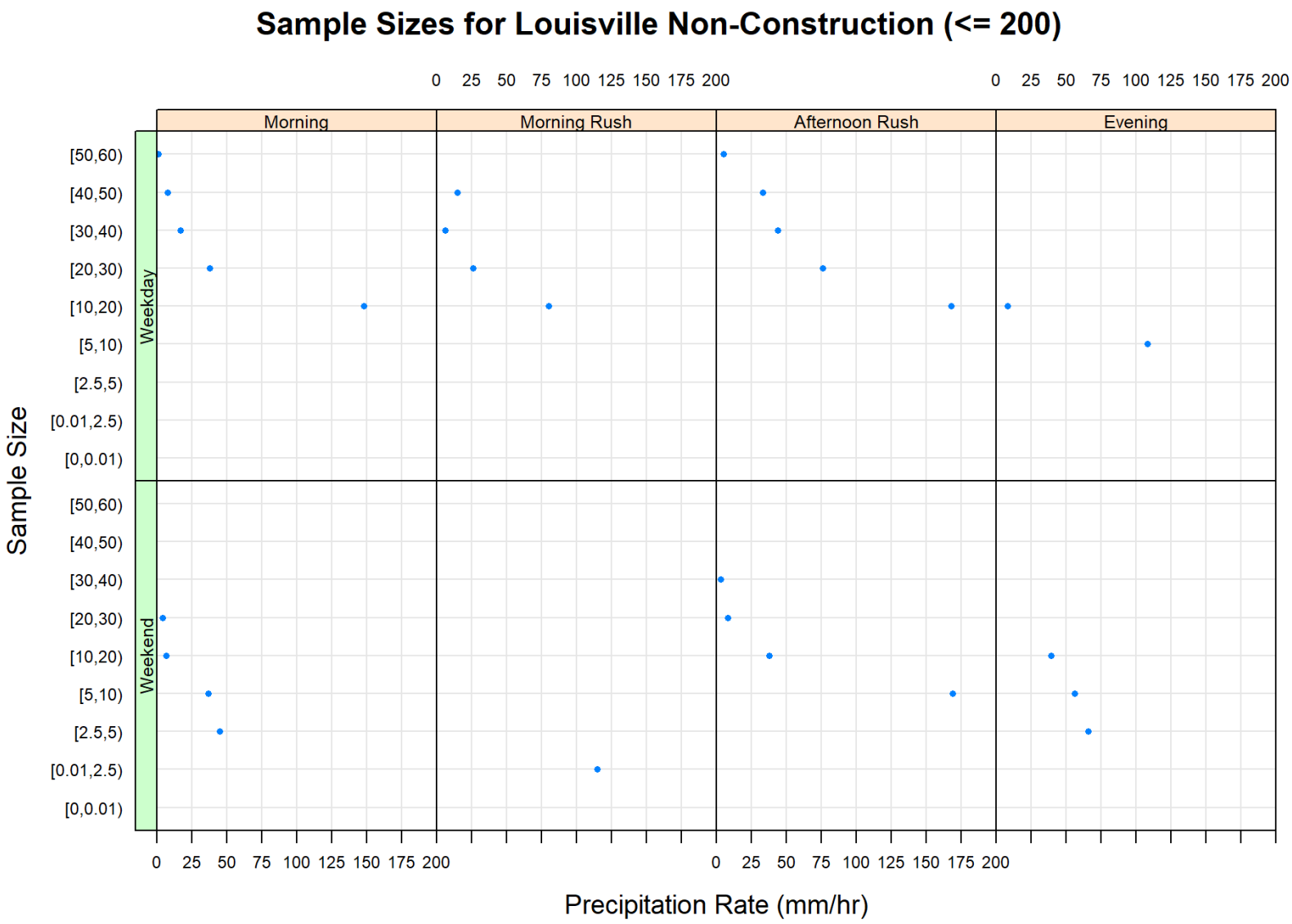


Figure 5.16. Sample size plots of Louisville for non-construction, hour ranges, and weekday/weekend

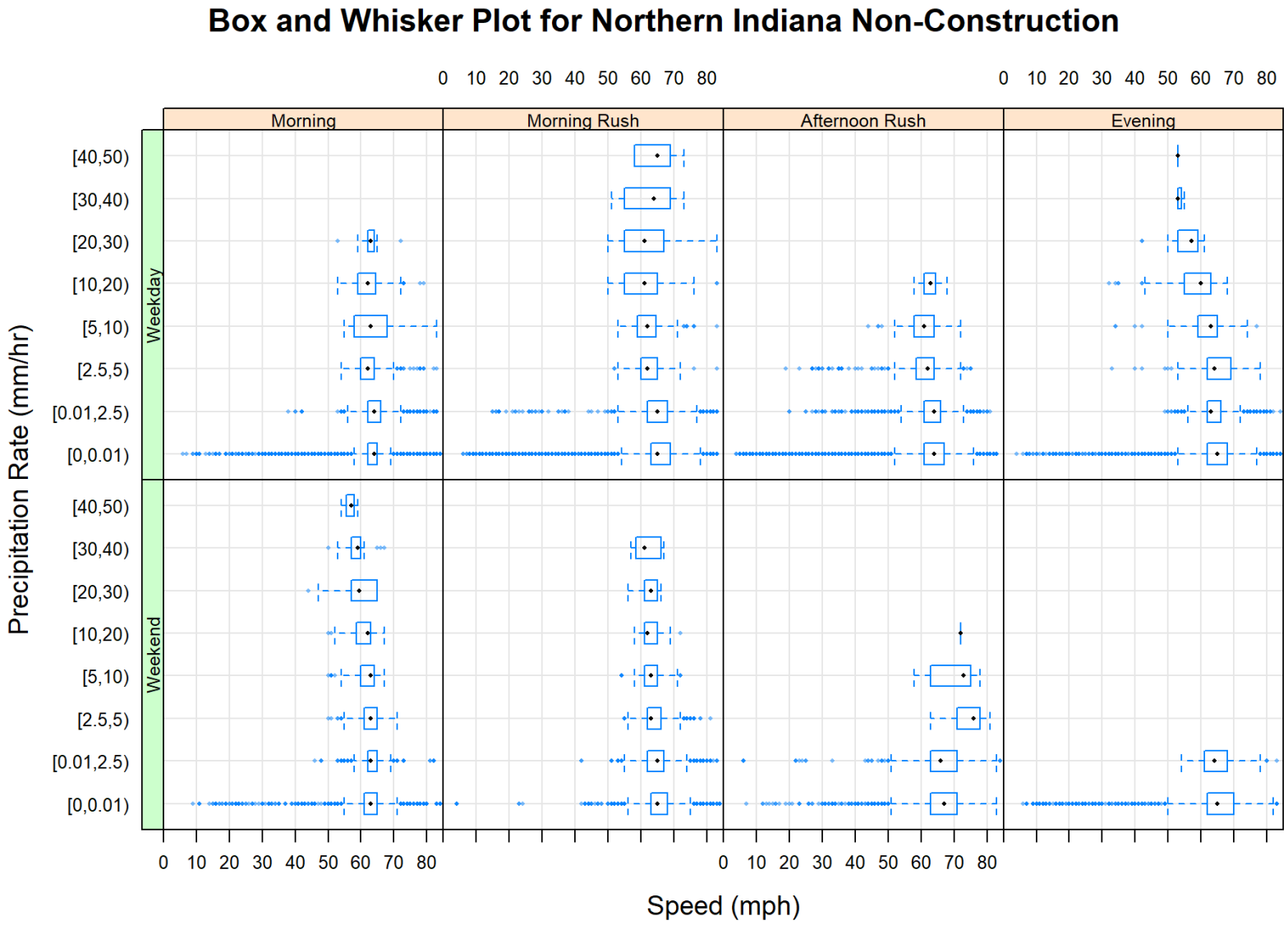


Figure 5.17. Boxplots of Northern Indiana for non-construction, hour ranges, and weekday/weekend

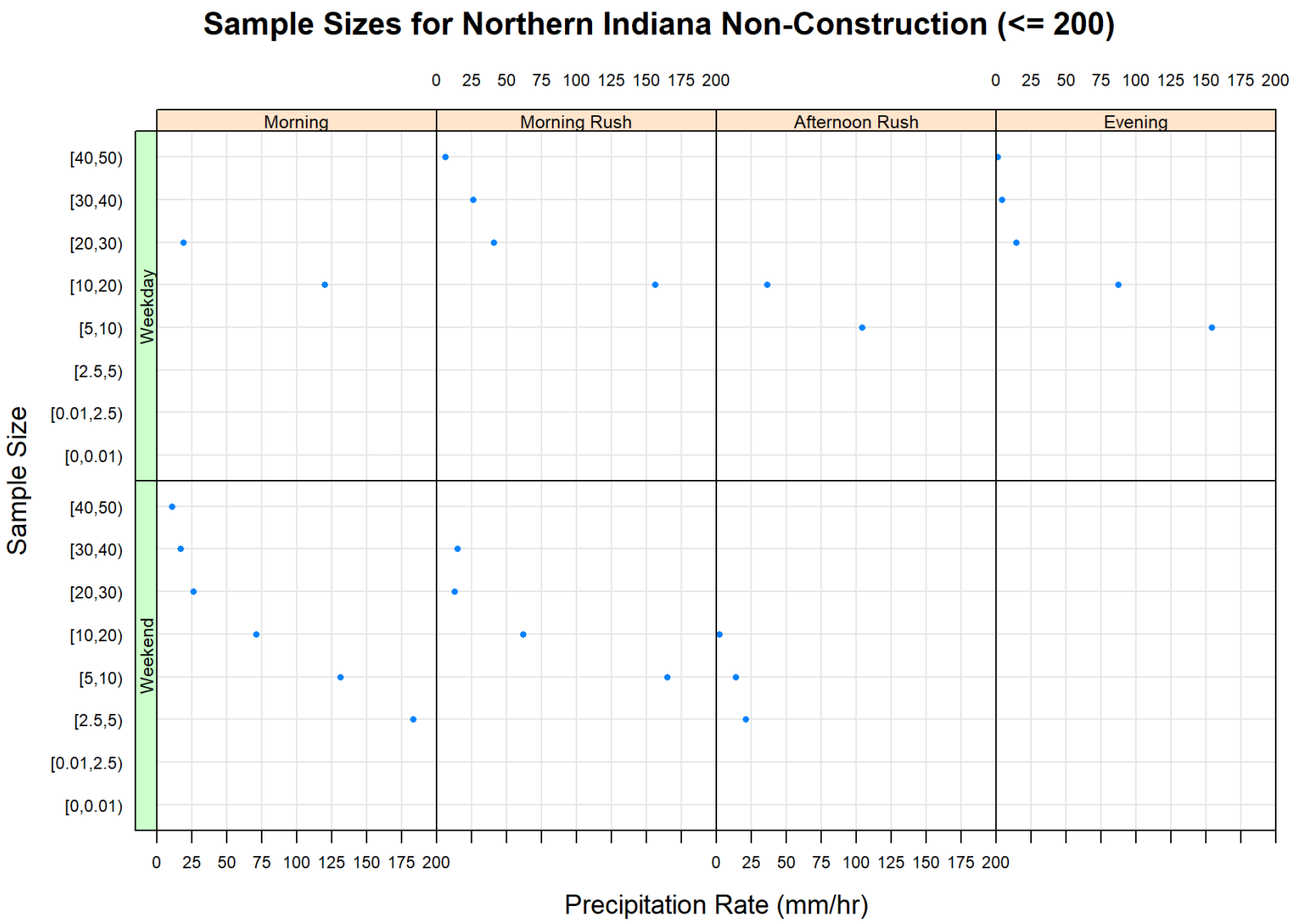


Figure 5.18. Sample size plots of Northern Indiana for non-construction, hour ranges, and weekday/weekend

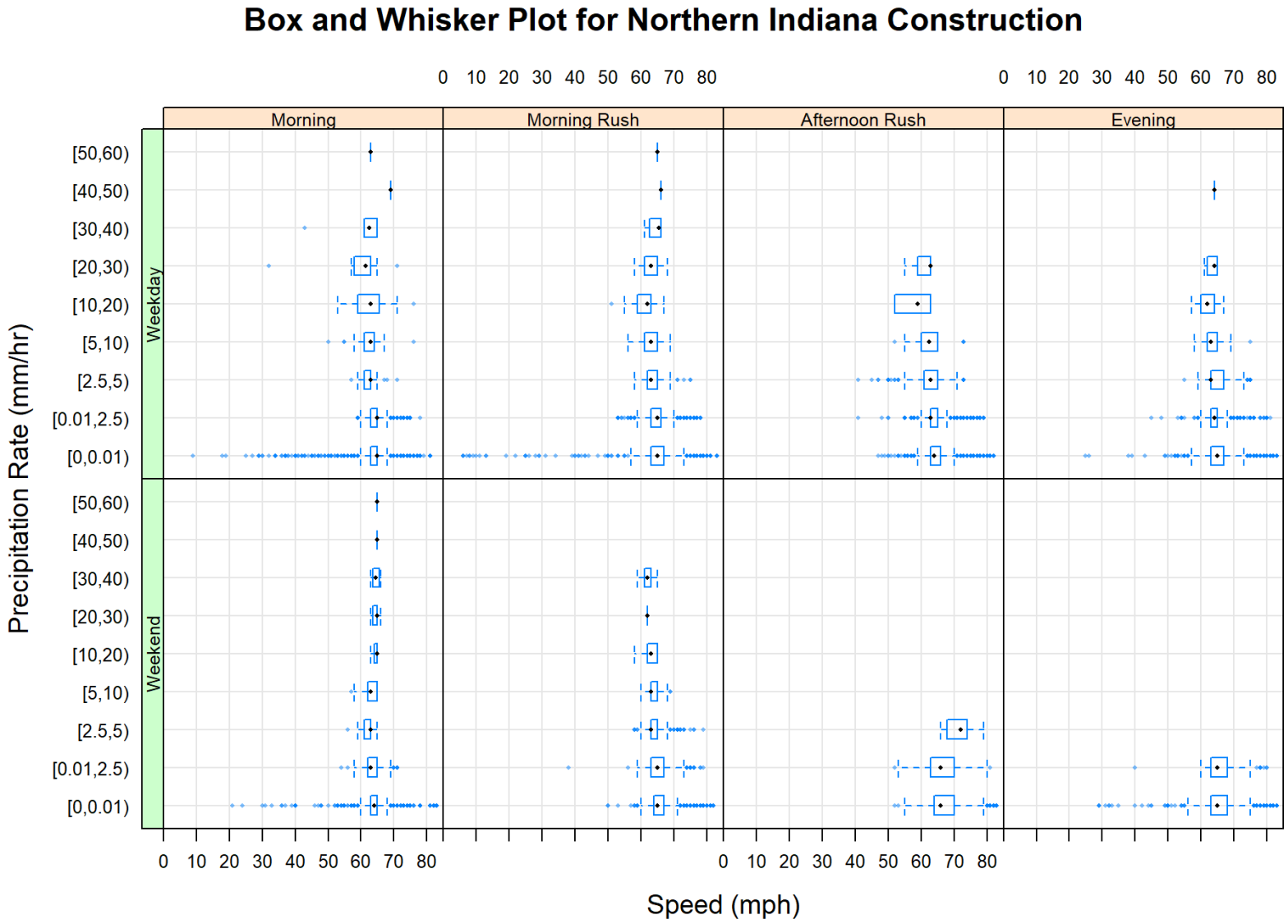


Figure 5.19. Boxplots of Northern Indiana for construction, hour ranges, and weekday/weekend

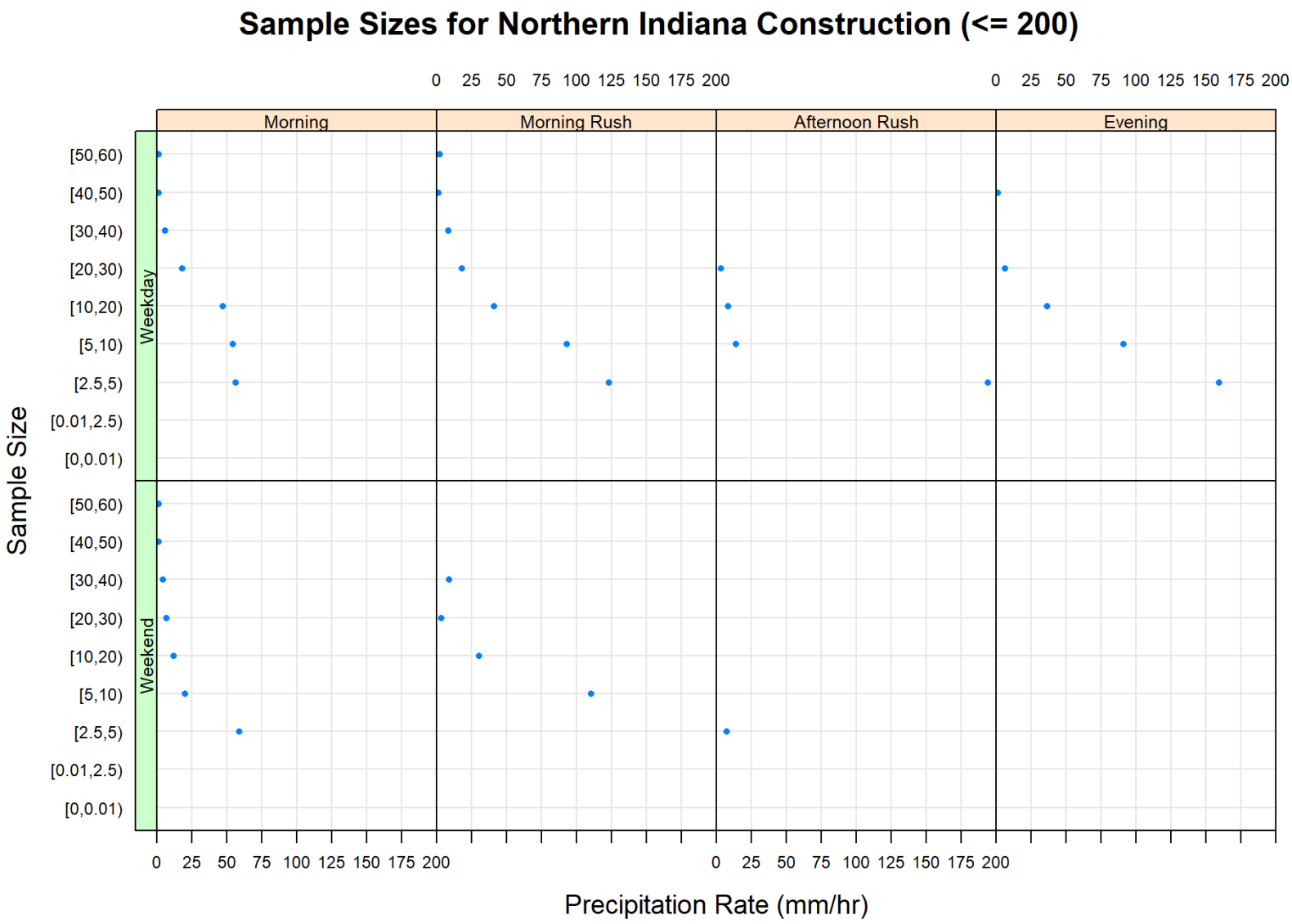


Figure 5.20. Sample size plots of Northern Indiana for construction, hour ranges, and weekday/weekend

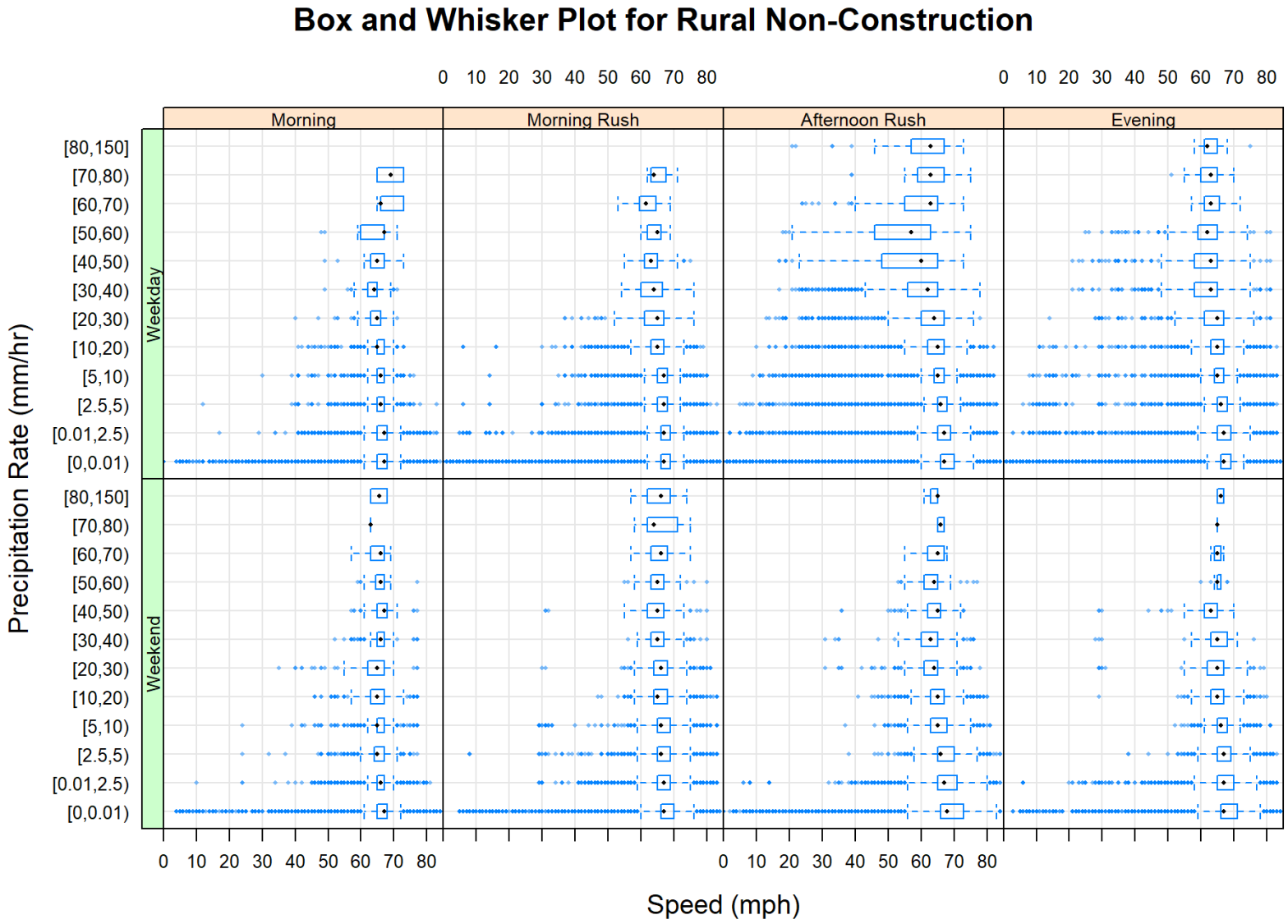


Figure 5.21. Boxplots of Rural areas for non-construction, hour ranges, and weekday/weekend

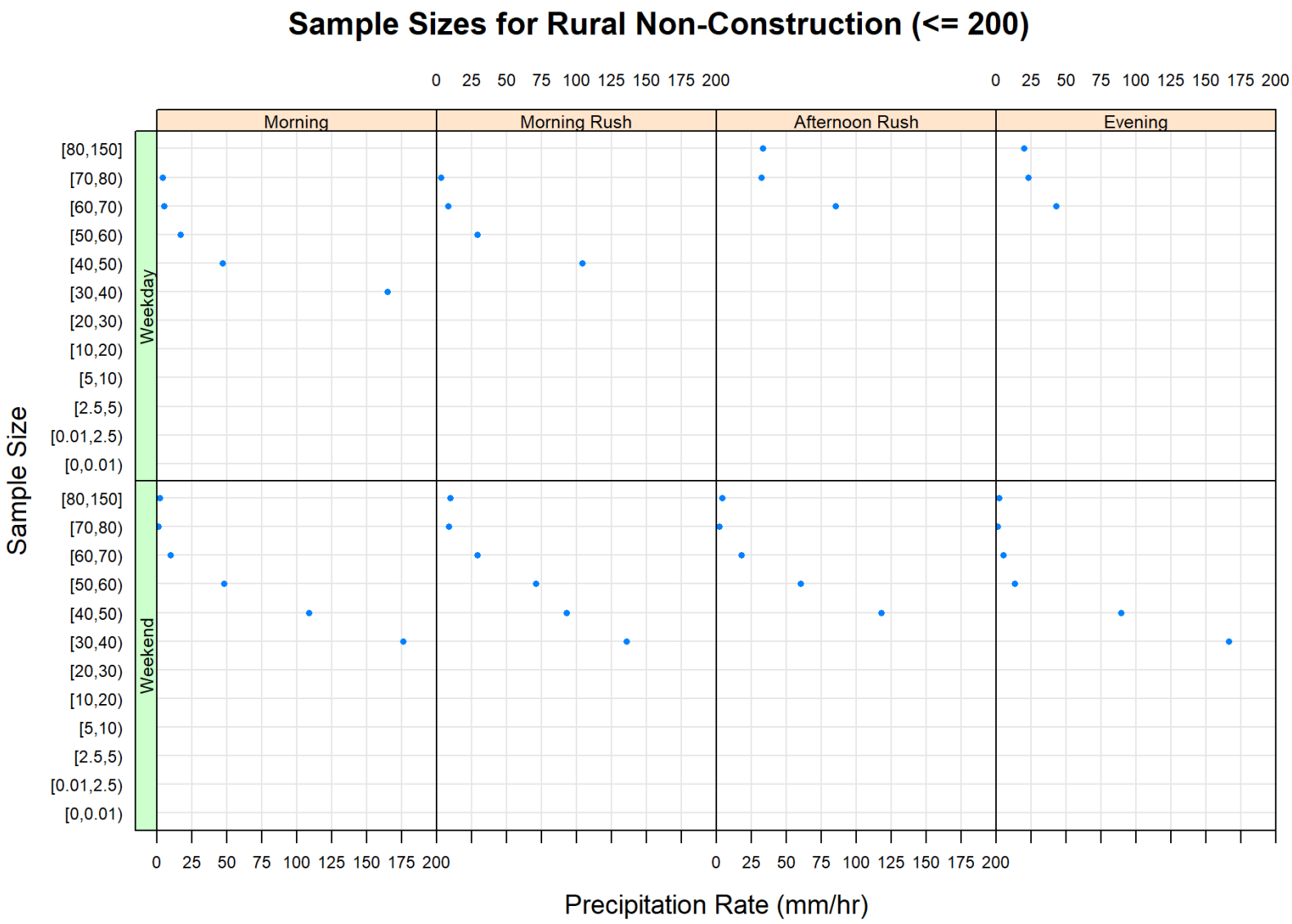


Figure 5.22. Sample size plots of Rural areas for non-construction, hour ranges, and weekday/weekend

Figure 5.23. Boxplots of Rural areas for construction, hour ranges, and weekday/weekend

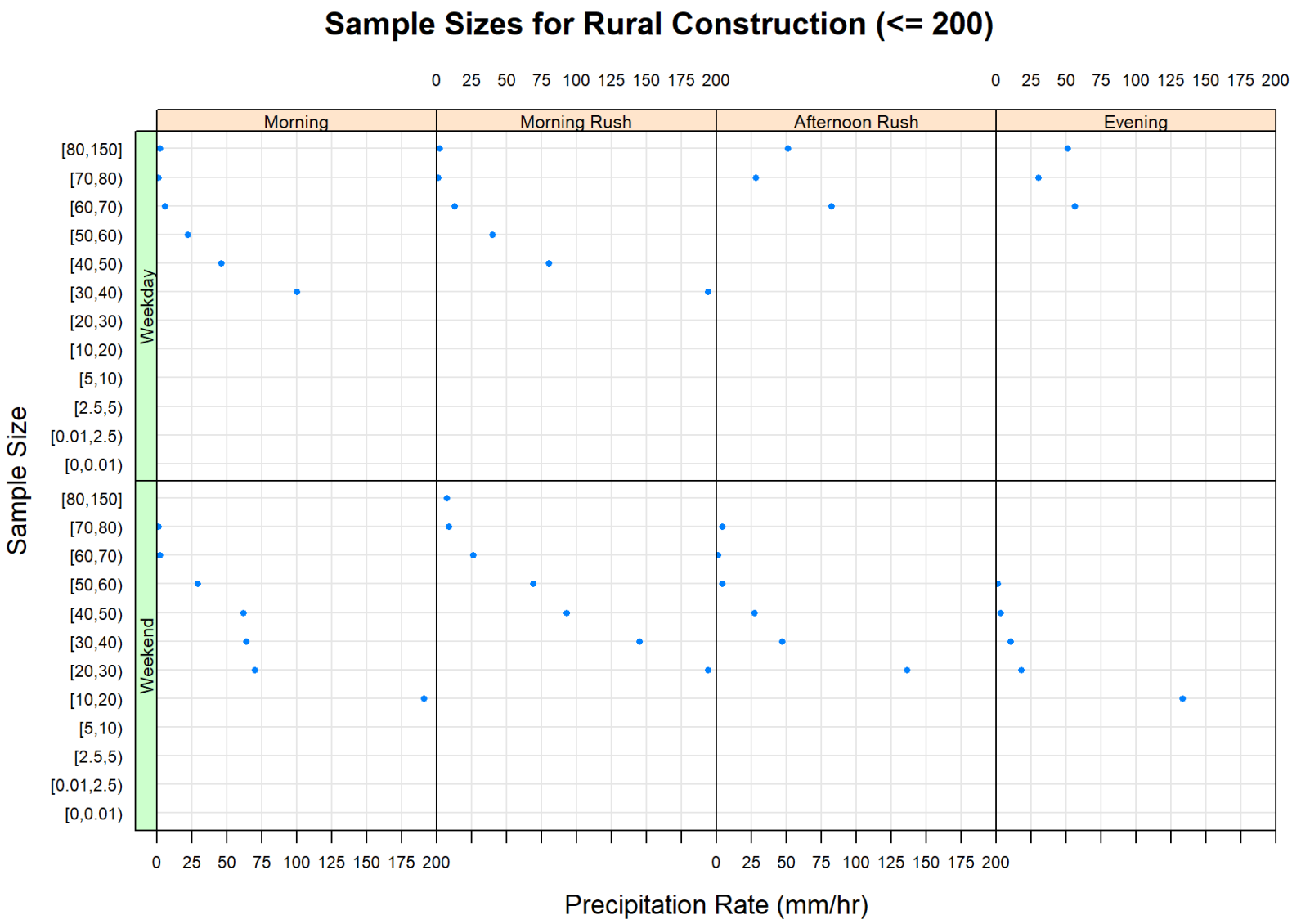


Figure 5.24. Sample size plots of Rural areas for construction, hour ranges, and weekday/weekend

The boxplots of traffic speeds under varying precipitation regimes provide a holistic look at the speed distributions, though a noisy one. To focus on the relationship between traffic speeds and rain, the median for non-rain scenarios is calculated as a baseline and used to calculate a delta, or reduction, for the median speed in rain scenarios. Each location has been plotted versus the precipitation regimes shown before, with the locations colorized for ease of comparison. The reduction plots can be seen in Figure 5.25 and Figure 5.26. The median was chosen as the primary measure of central tendency here for its resistance to outliers. When reviewing the speed reductions, it is important to keep in mind the variability that was demonstrated in the boxplots from Figure 5.13 through Figure 5.23. Additionally, not all scenarios have a sufficient sample size for analysis. Thus, only scenarios in which the sample size is greater than 100 have been plotted.

An initial review of the non-construction speed reductions in Figure 5.25 shows a quasi-linear trend towards increased speed reductions from the non-rain median speed. The morning rush and afternoon rush time frames display the strongest linear trend in each location. In the case of the afternoon rush scenario, the speed reductions are much more dramatic during the weekday than during the weekend. This is a somewhat intuitive result as traffic volumes are higher during these time frames. The evening time frame also seems to exhibit this linear trend of higher speed reductions under higher precipitation regimes. The morning time frame seems fairly insensitive to precipitation intensity as the speed reductions are either 0 or are constant as precipitation intensity increases. One particular point of interest is the positive speed reductions during the morning weekday time frame. A positive speed reduction would indicate that the median speeds during rain scenarios are faster than the median speeds in non-rain scenarios. While this is counterintuitive, the actual speed difference is not that significant, 2—3 mph, and occurs only during a low volume time frame. The largest speed reduction shown here is 10 mph in rural regions during the afternoon rush timeframe on a weekday.

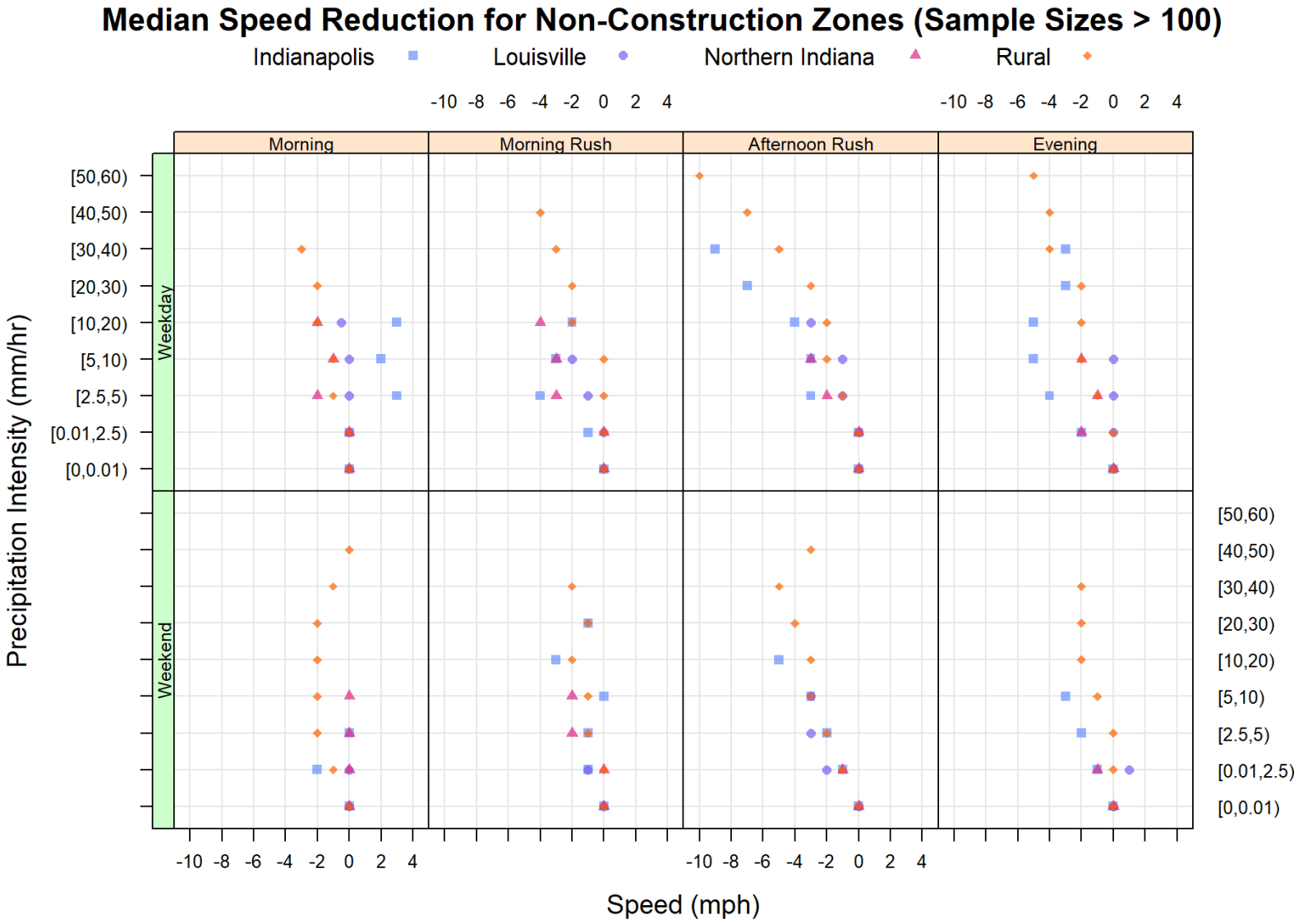


Figure 5.25. Median speed reduction in non-construction zones

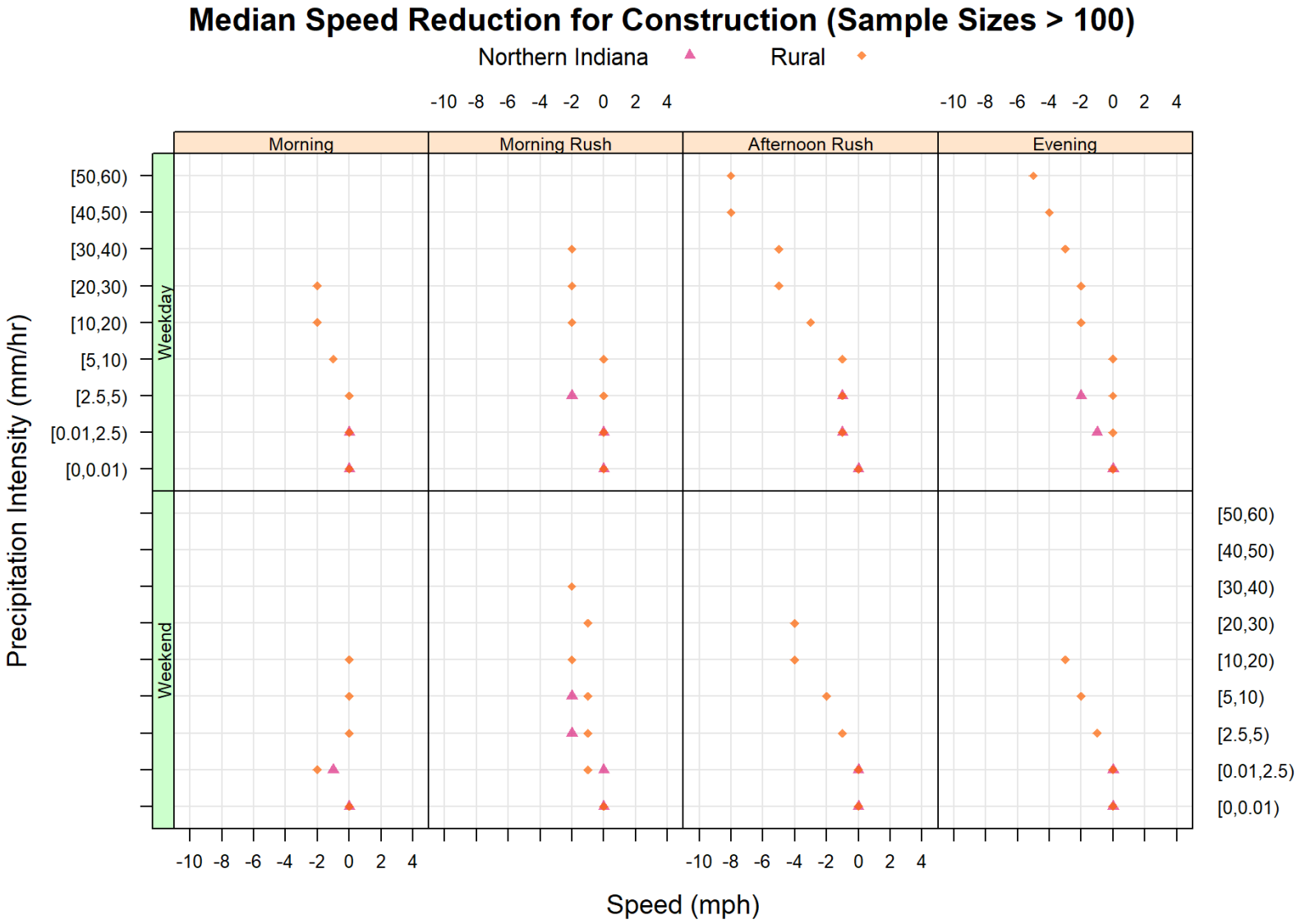


Figure 5.26. Median speed reduction in construction zones

Figure 5.26 shows the speed reductions for each time frame and location in the presence of construction. Indianapolis and Louisville are not plotted here as these locations did not have any construction for the month of June 2018. The conclusions drawn from Figure 5.25 seem to be comparable here. The morning time frame seems relatively insensitive to higher precipitation intensities while the morning rush, afternoon rush, and evening time frames seem to be much more sensitive. There does not seem to be an appreciable difference between the non-construction and construction zones. It would seem that the speed reductions are comparable in both situations.

5.2 Speed Prediction

Given that XGBoost is well renowned for working quite well on weak learners, it was hypothesized that it could be useful in the domain of traffic speed studies with exogenous variables like precipitation intensity included. As there was little literature regarding this method for this particular domain, it was decided that it would be used by this study. The results were particularly poor despite attempts at improving the model by varying the eta and boosting iteration parameters, as seen in Figure 5.27. The MAE values for each combination of region, construction/non-construction, hour range, and weekday/weekend were calculated and can be seen in Figure 5.28 and Figure 5.29.

The XGBoost model, despite having an autoregressive term, failed to produce tenable results capable of even beating a 60-minute moving average, let alone a 10-minute moving average. This does not mean it failed completely though. Review of Figure 5.28 shows that XGBoost was relatively close to beating the 60-minute moving average, particularly in Louisville. This suggests that perhaps with more comprehensive feature engineering, XGBoost could potentially yield substantially better results. The 3-fold cross validation in Figure 5.27 seems to indicate other hyperparameters as a potential place to begin further tuning, aside from further feature engineering. Comparison of Figure 5.28 and Figure 5.29 does not seem to indicate any appreciable difference in errors between construction and non-construction zones.

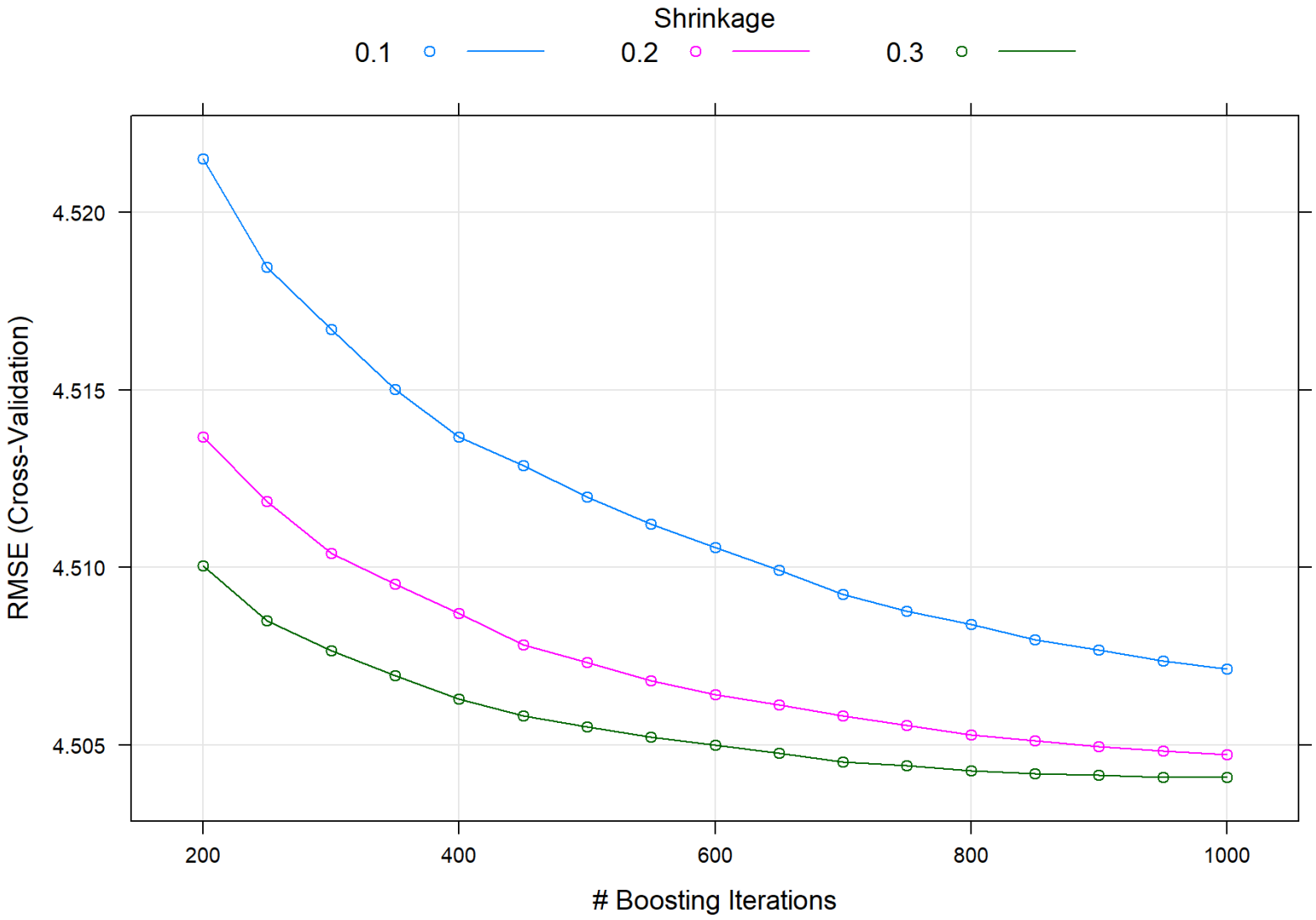


Figure 5.27. Results of 3-fold cross validation for XGBoost model with eta and boosting iterations varied

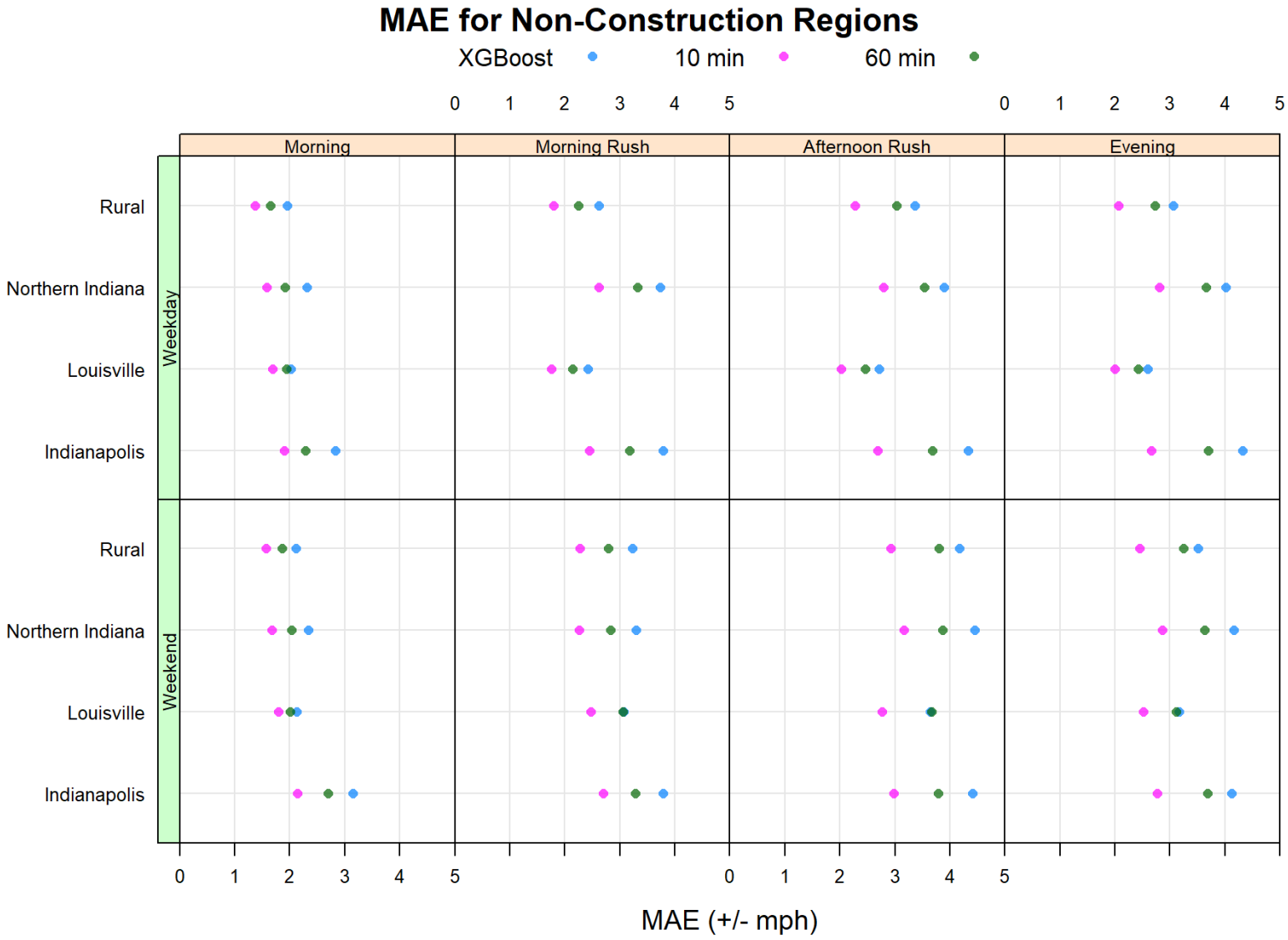


Figure 5.28. MAE values calculated for Non-Construction regions by weekday/weekend and hour range

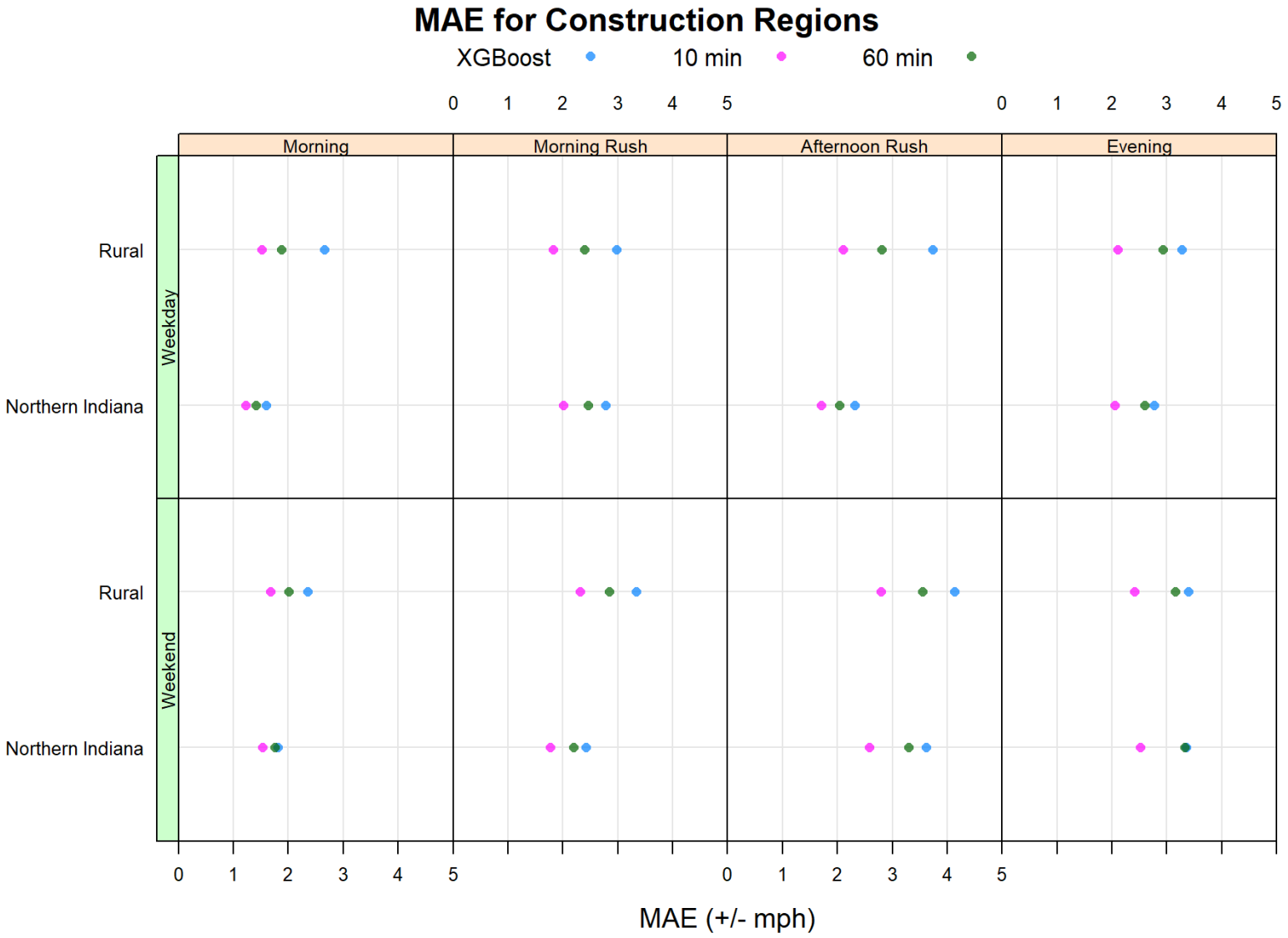


Figure 5.29. MAE values calculated for construction regions by weekday/weekend and hour range

CHAPTER 6. CONCLUSIONS

This study investigated the interaction between interstate 65 traffic speeds in Indiana and precipitation intensity for June 2018. In order to do this, the speed distributions were investigated through comparison of rain and non-rain traffic speeds under varying conditions such as weekday/weekend, hour range, construction/non-construction, and region. It was found that generally, above the 5^{th} percentile, traffic speeds were faster or comparable in non-rain scenarios as opposed to rain scenarios. Below the 5^{th} percentile, traffic speeds tended to be faster in non-rain scenarios. It was noted that these visualizations do not account for the variability in the data and that the faster rain speeds were a very small proportion of the dataset.

The rain/non-rain traffic speeds were compared with a gamma distribution in order to determine if the speeds followed some form of a standard distribution. Traffic speeds were limited to 50 mph and above in an attempt to produce a better fit and limit any outliers. The results showed that non-construction zones, with the exception of Louisville, seemed to be fit by the gamma distribution better than construction zones across regions. As precipitation intensity was considered via boxplots, it was shown that there is a significant amount of variability in traffic speeds across all precipitation regimes. Due to this variability, a median non-rain traffic speed baseline was established to draw comparisons against. The results indicated that there is a quasi-linear relationship, as precipitation intensity increases, traffic speeds tend to fall. The amount of speed decreases seems to be region, weekday/weekend, and hour range dependent. The most significant traffic speed slowdowns were shown to occur during the afternoon rush on a weekday with Rural areas showing the greatest speed decrease of 10 mph at a 50-60 mm/hr precipitation intensity.

Finally, an XGBoost model was developed to see if traffic speeds could be predicted better than a simple naïve forecast of traffic speeds. The model was relatively simple with the speed as the response variable and a 16-minute lagged speed, precipitation intensity, region, bearing, construction/non-construction, hour range, day/night, weekend/weekday, and day of the week as predictors. The model did not perform particularly well and was easily outperformed by a 10-minute moving average as well as a 60-minute moving average. Given the closeness of the MAE values from the model, it does seem to suggest that further feature engineering and hyperparameter tuning could result in a more accurate model.

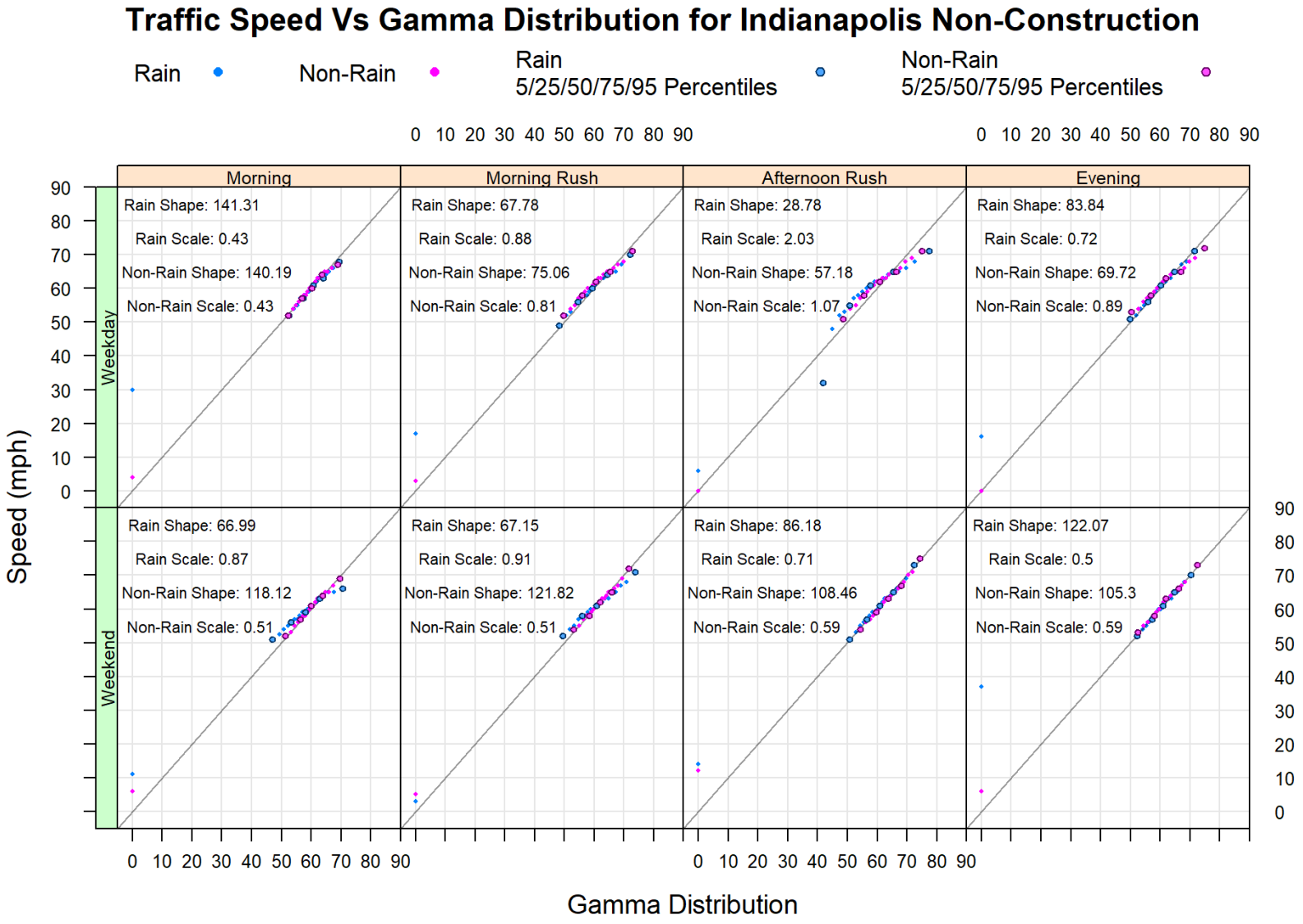
The key findings from this research are as follows:

1. Non-rain traffic speeds above the 5th percentile in the month of June are typically faster than their rain speed counterparts.
2. There is a significant amount of variability in the traffic speeds under various precipitation intensities.
3. A gamma distribution does not adequately fit traffic speeds under rain and non-rain scenarios in all situations and cannot be formulated in a general fashion.
4. The afternoon rush is the most strongly impacted time frame during rain events seeing speed reductions of up to 10 mph.
5. XGBoost does not perform adequately for speed predictions, as done in this study.

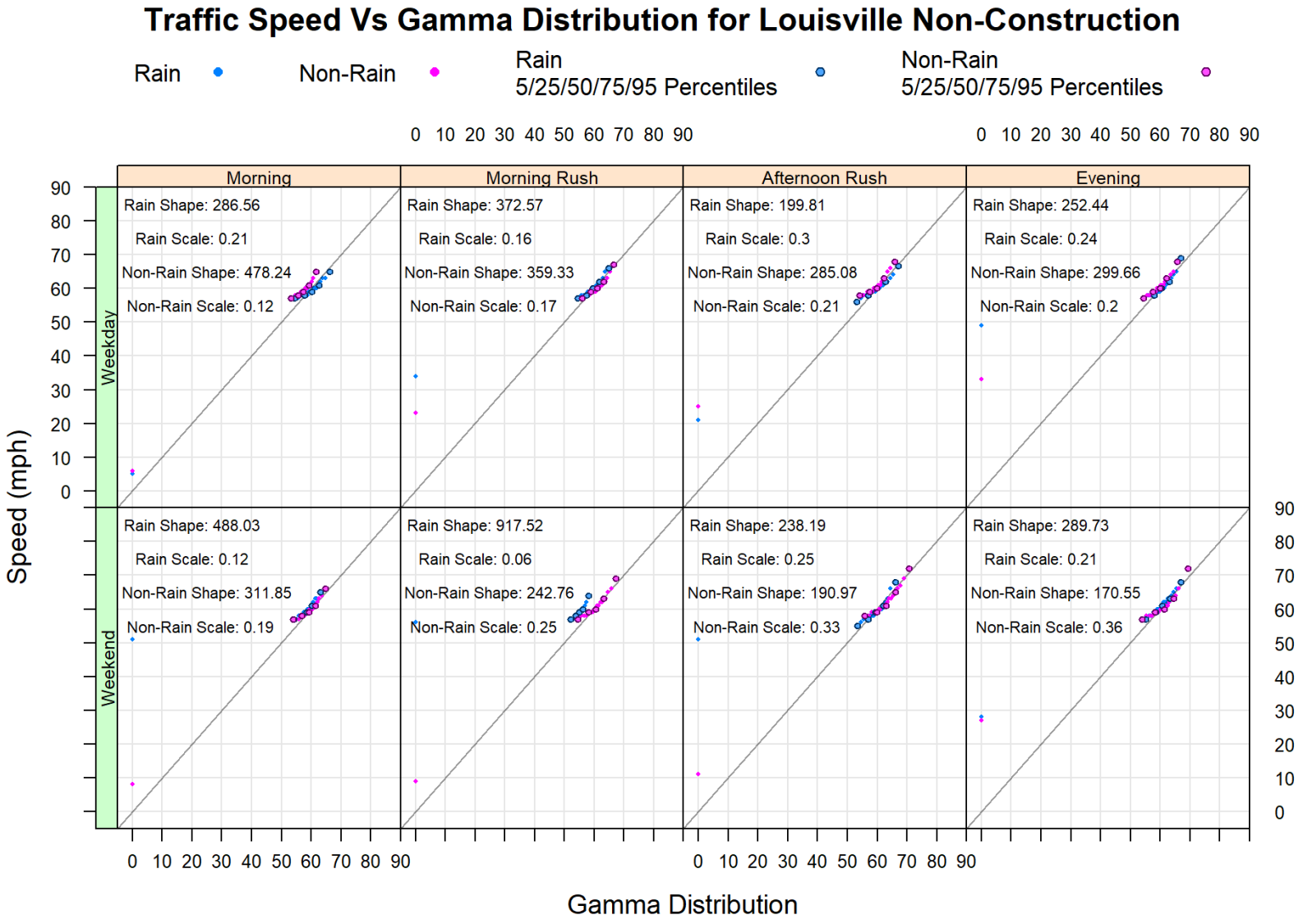
6.1 Future Work

For future work, this study has two recommendations. The first recommendation is to investigate the use of high-resolution weather models for furthering this kind of research. The High Resolution Rapid Refresh (HRRR) model would seem to be an excellent candidate for traffic studies. It has a 15-minute temporal resolution, 3-by-3 square-kilometer spatial resolution, and 55 variables available at the surface, 2-meter, and 10-meter levels. No archives of the 15-minute data, which would be necessary for fine-scale traffic interaction studies, currently exist. Thus, this study recommends that future researchers begin archival of the 15-minute HRRR forecast data, at least at the 0th hour analysis step, in order begin to relate many weather variables at a high resolution to traffic speeds. This may prove particularly useful for winter operations. The second recommendation this study makes is to perform future analyses of weather and traffic speeds at the 15-minute level. Finer time resolutions may be appropriate in some cases, but the 15-minute level should help to reduce noise in the data as well as align studies with the temporal resolution of high-resolution weather datasets like the HRRR.

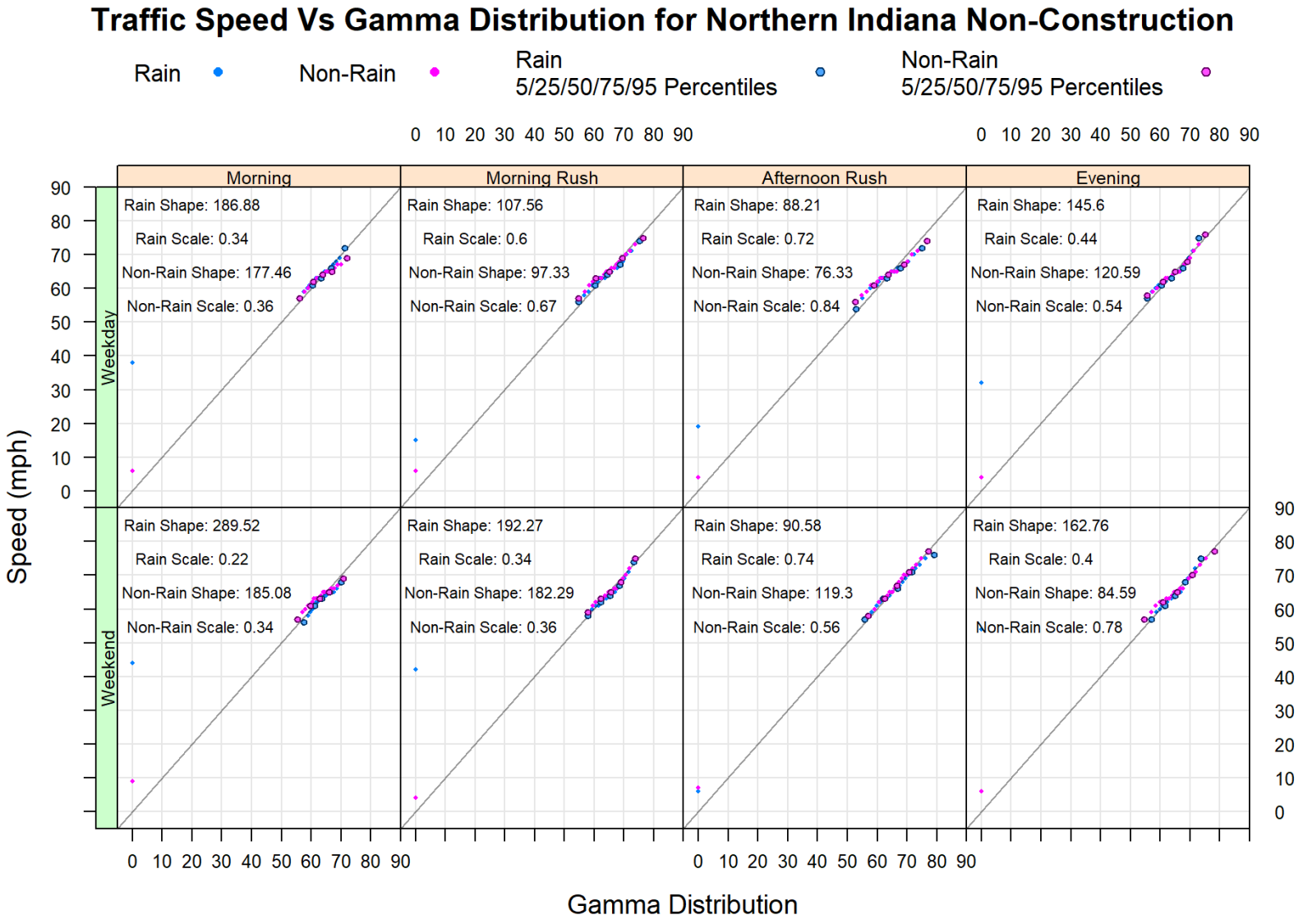
APPENDIX A. SUPPLEMENTAL FIGURES



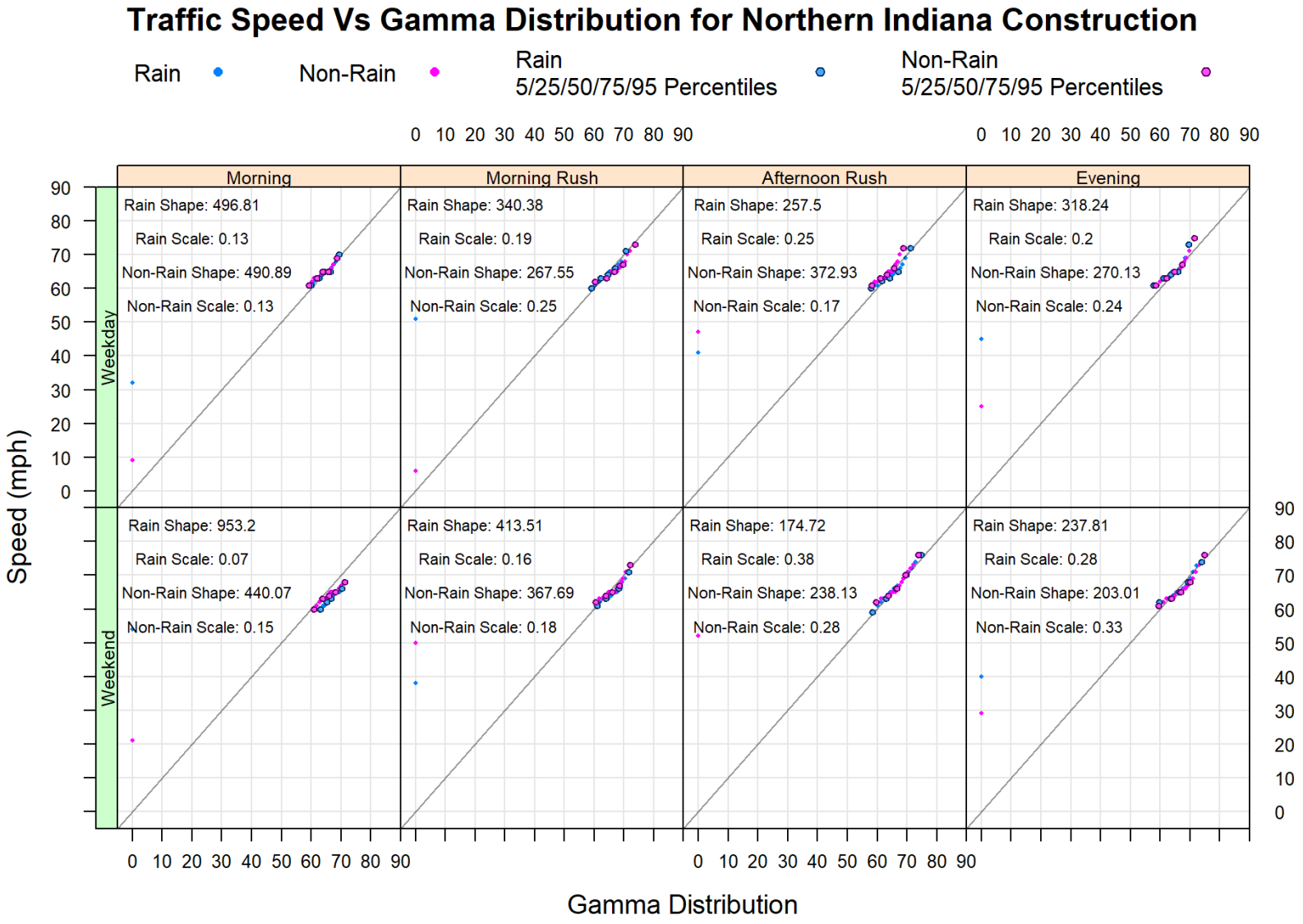
Appendix A 1. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Indianapolis, non-construction



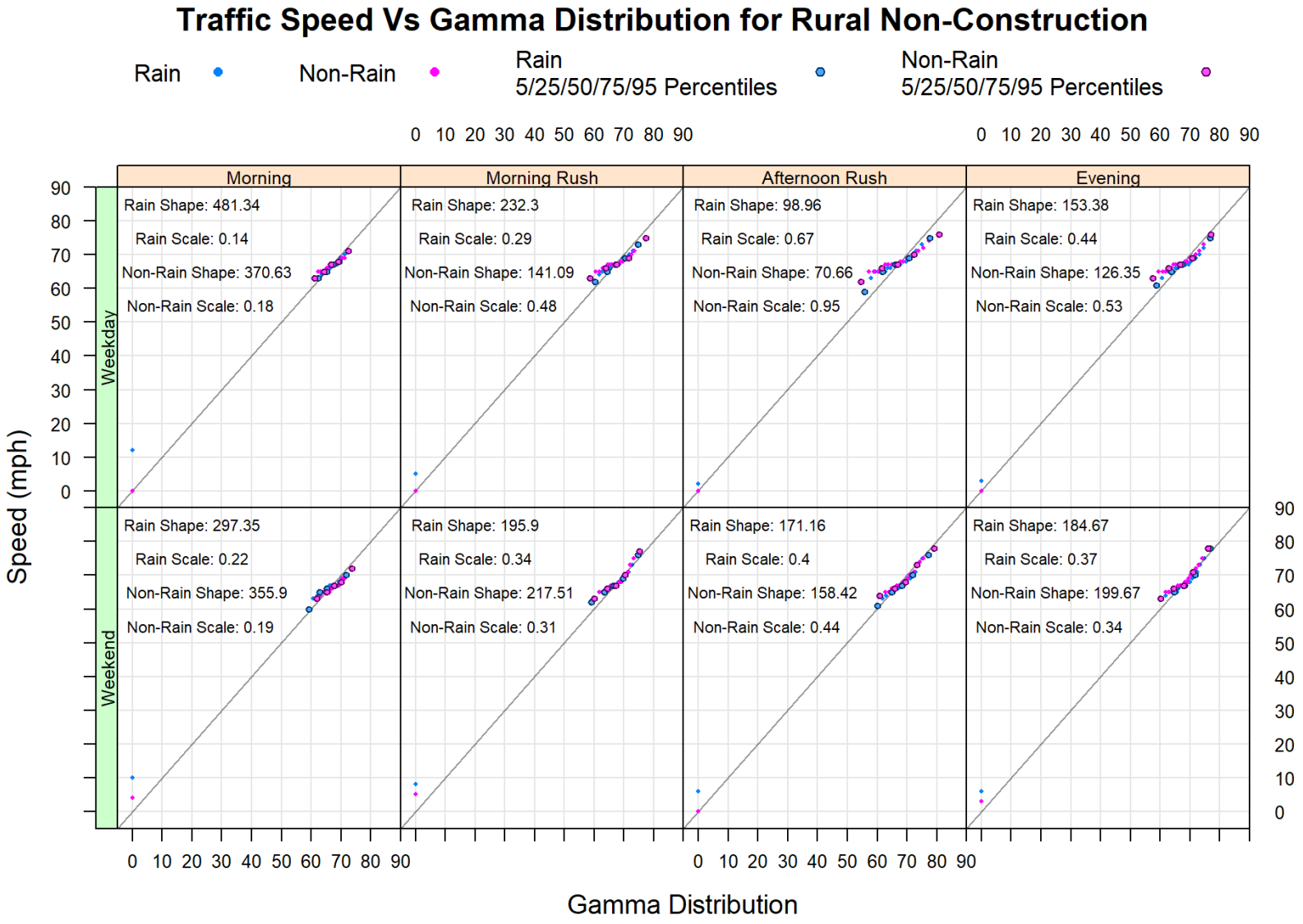
Appendix A 2. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Louisville, non-construction



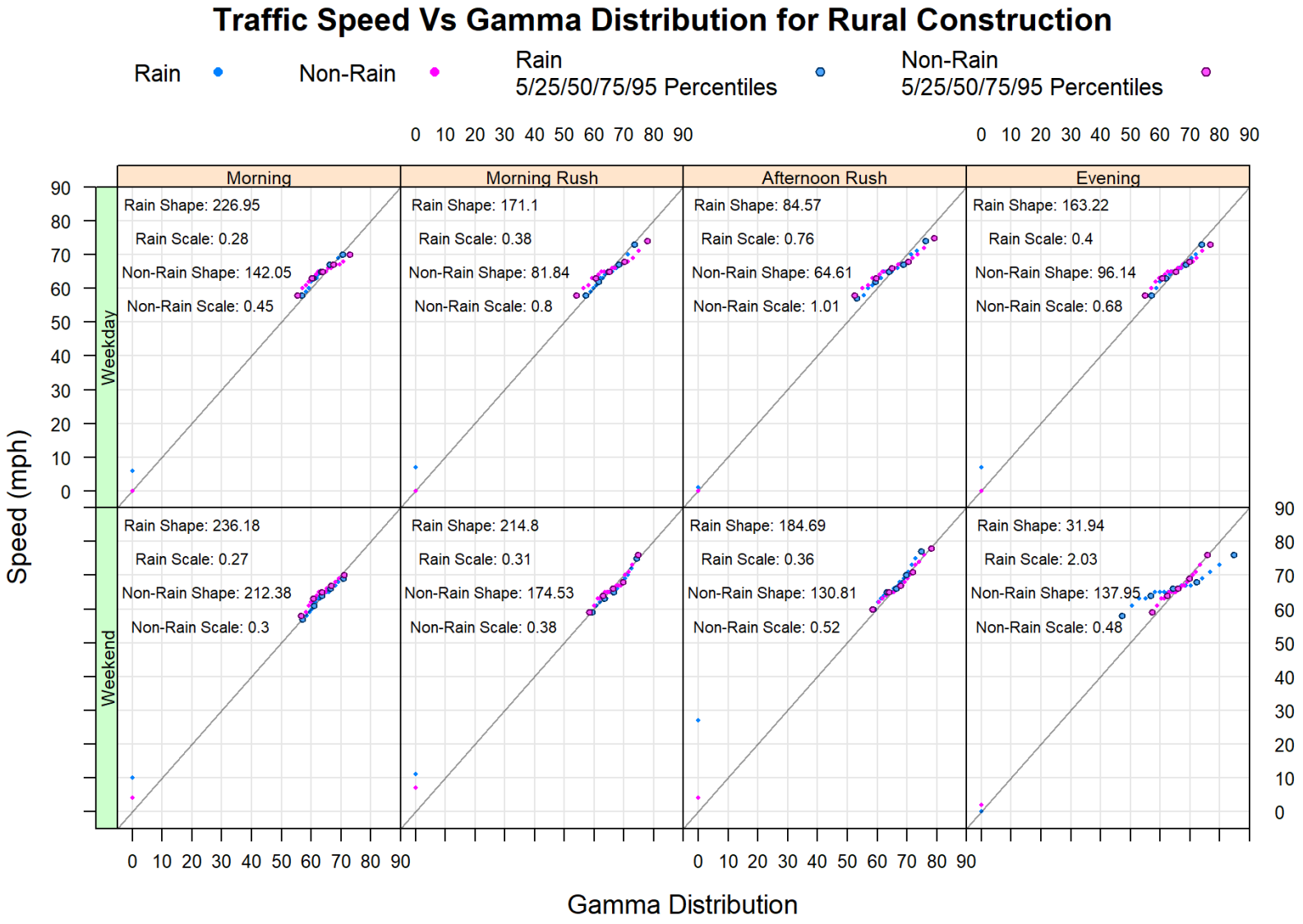
Appendix A 3. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Northern Indiana, non-construction



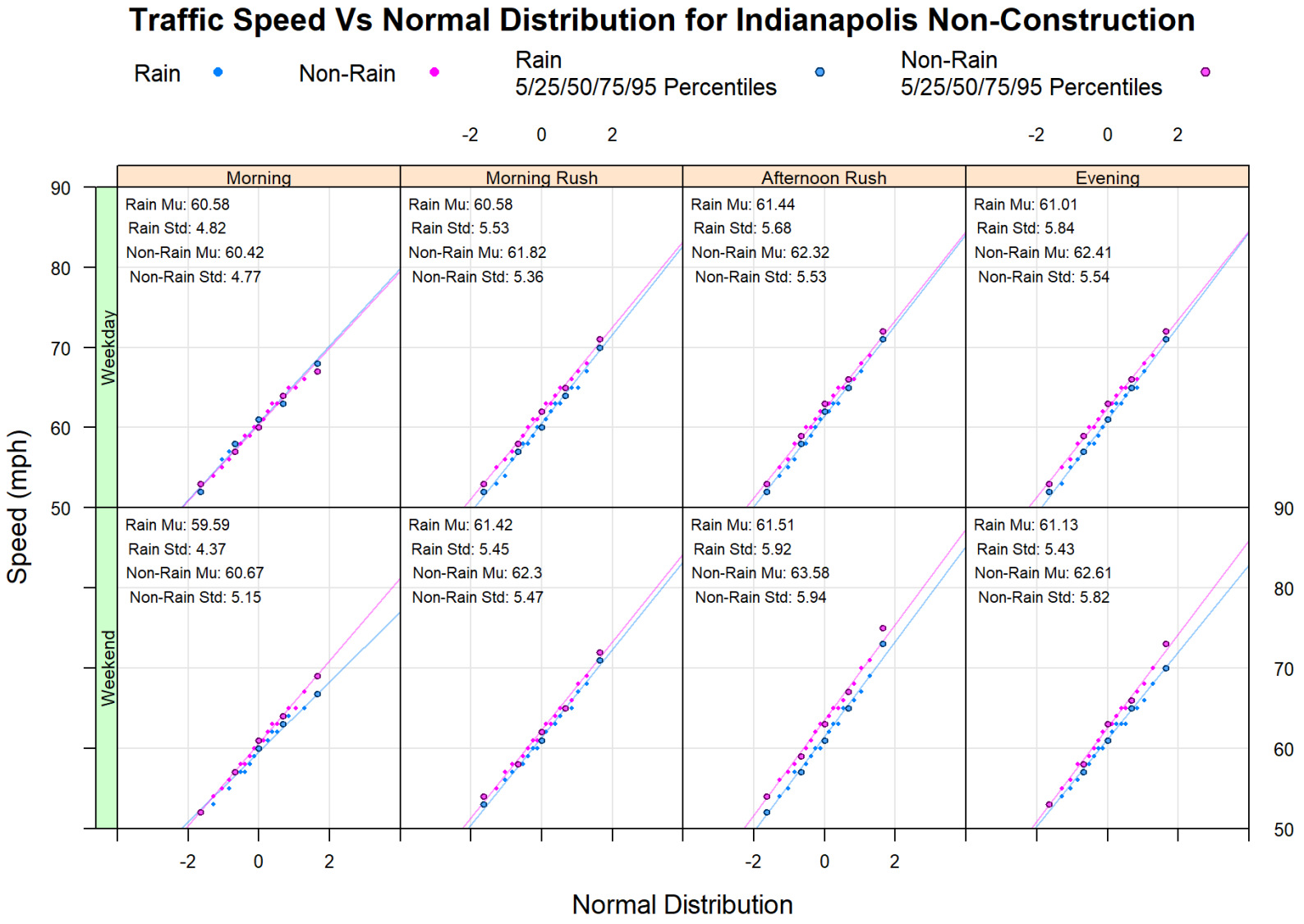
Appendix A 4. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Northern Indiana, construction



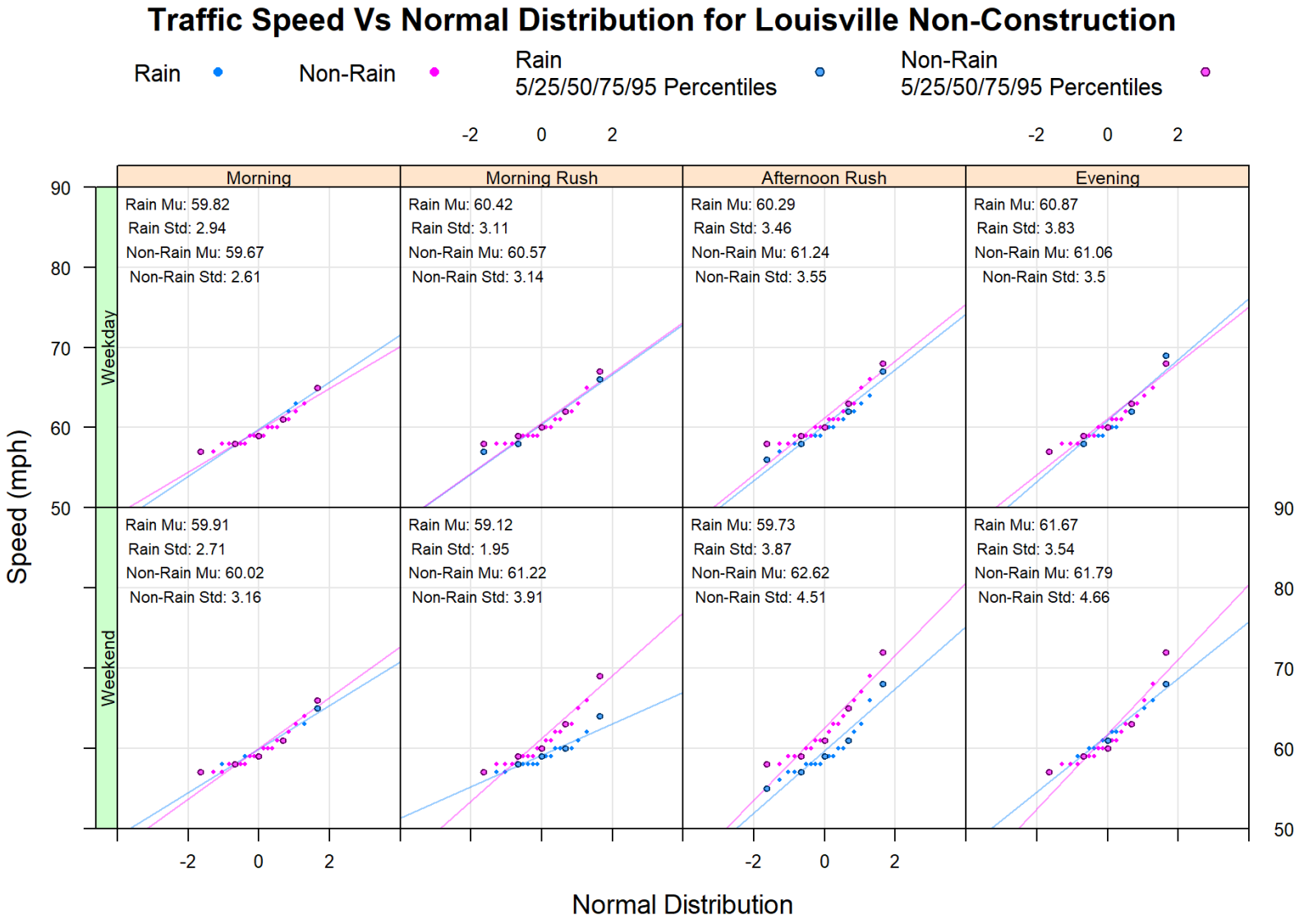
Appendix A 5. . Gamma distribution fits of rain and non-rain traffic speeds (mph) for Rural areas, non-construction



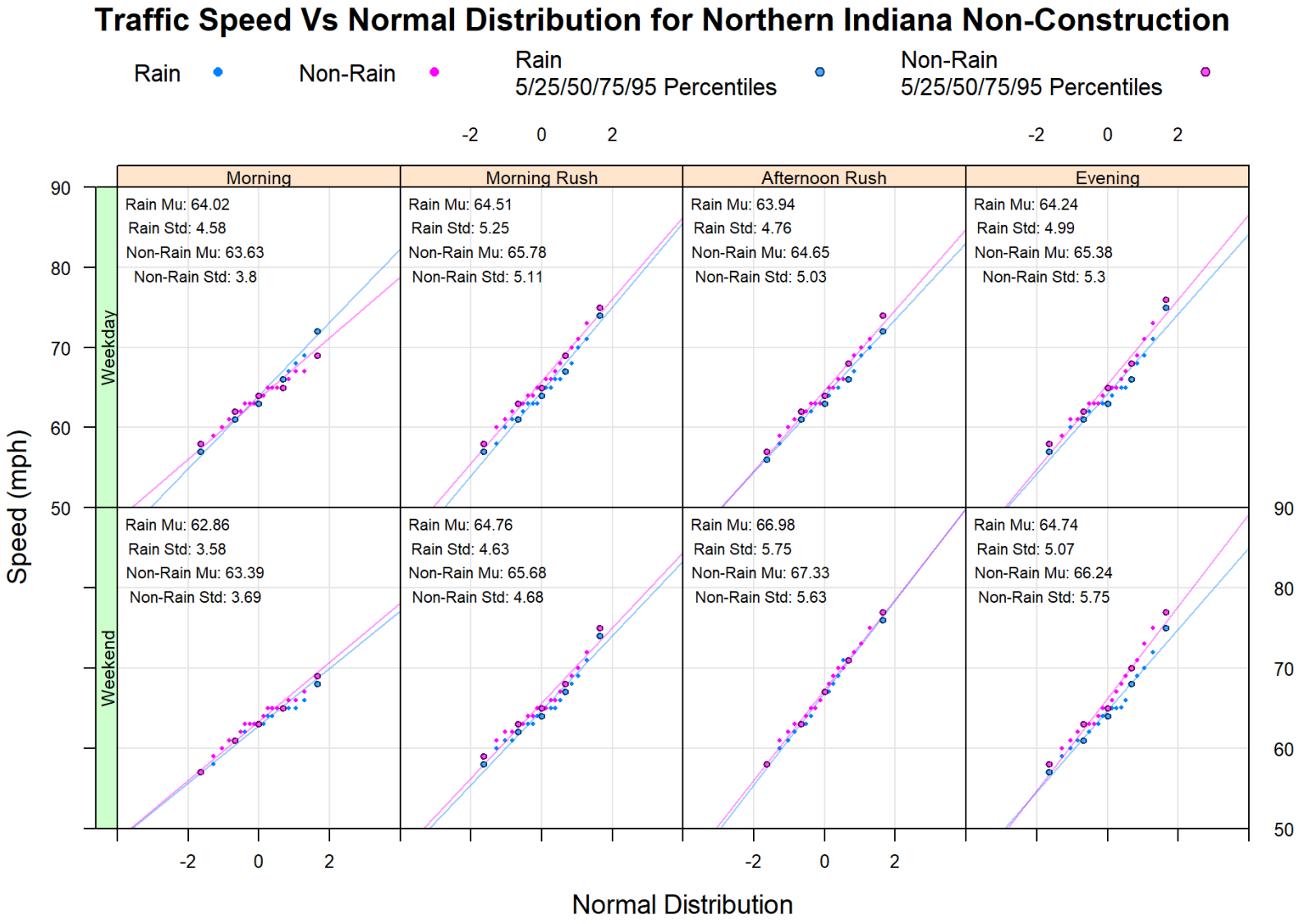
Appendix A 6. Gamma distribution fits of rain and non-rain traffic speeds (mph) for Rural areas, non-construction



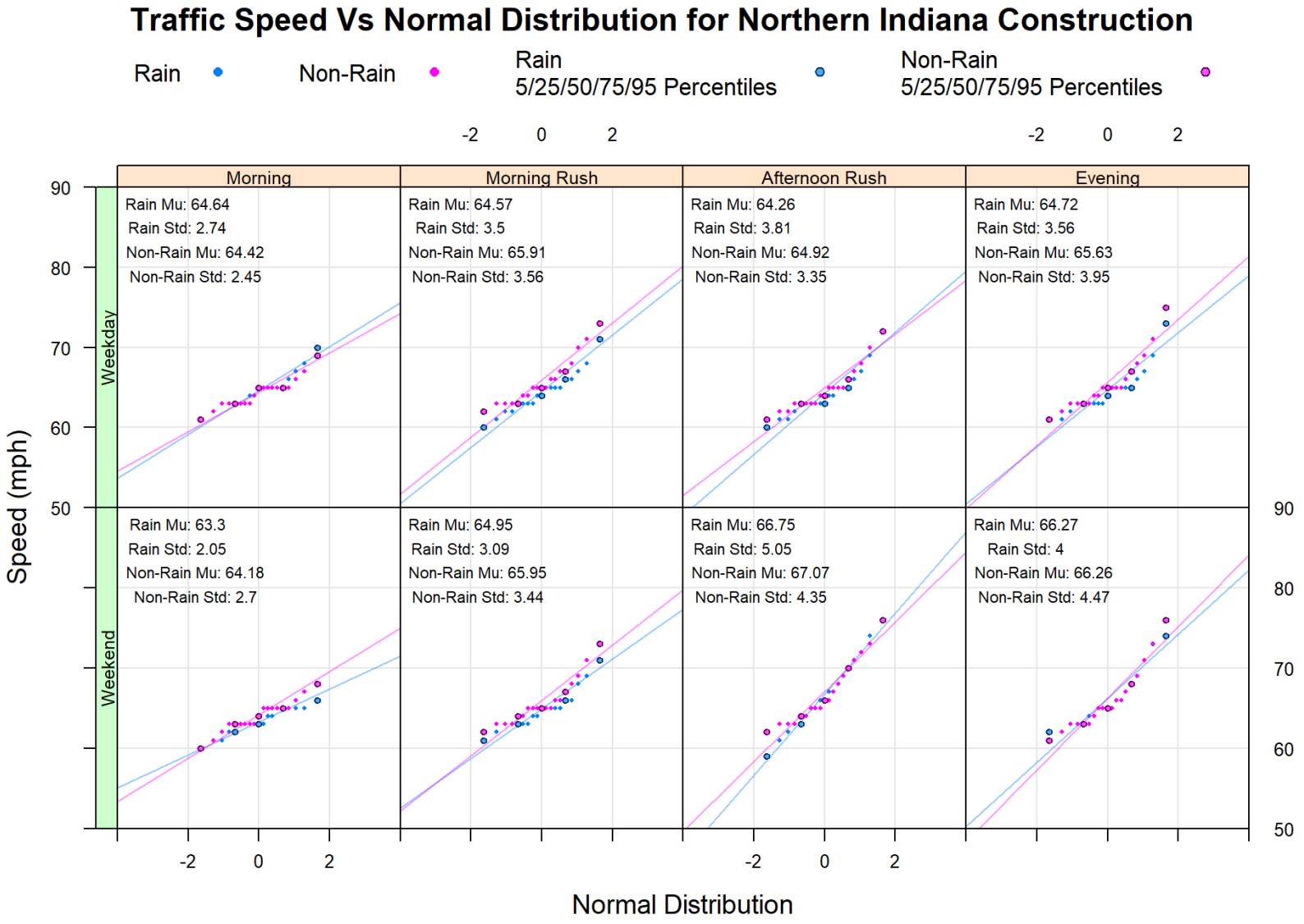
Appendix A 7 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Indianapolis, non-construction



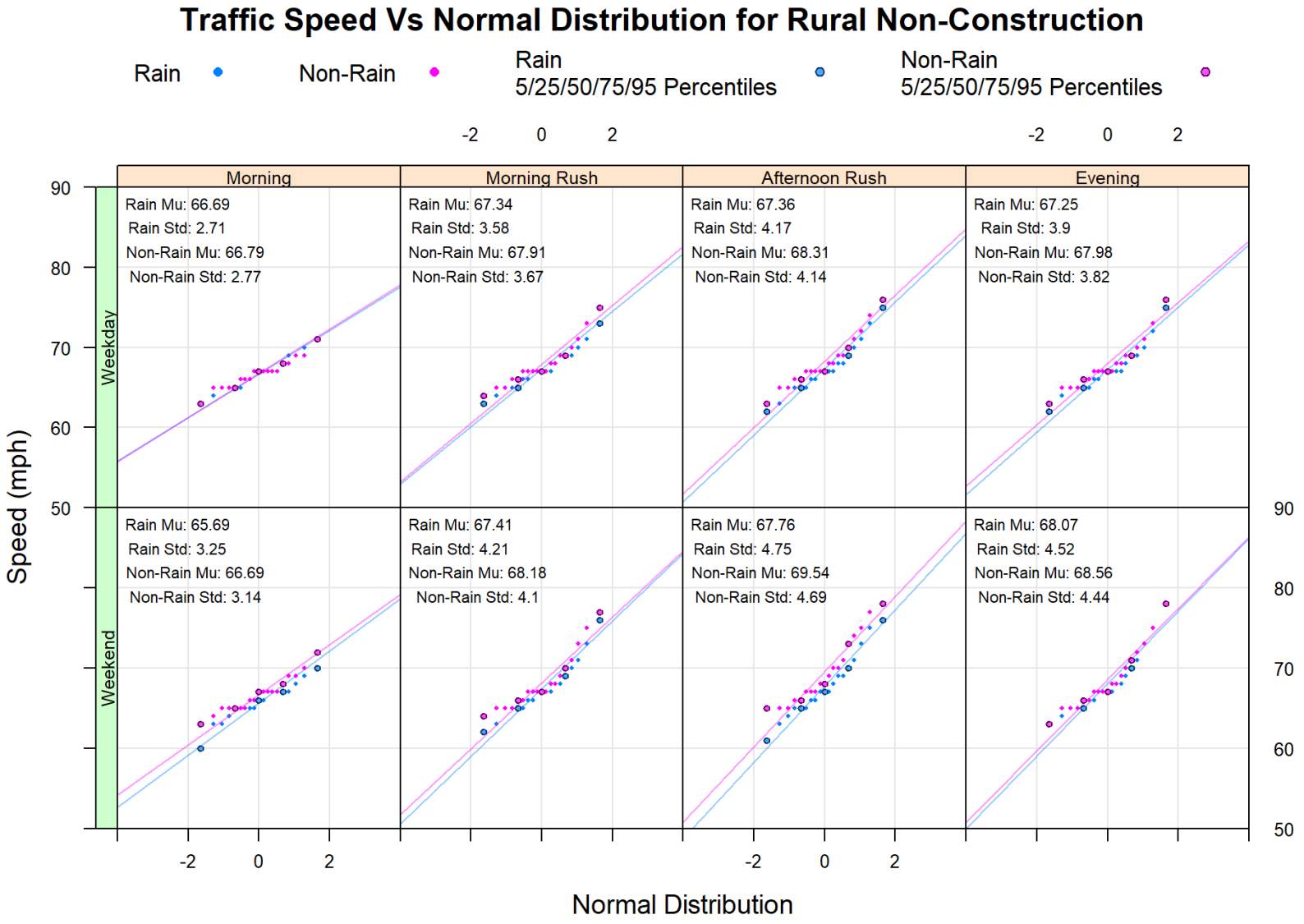
Appendix A 8 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Louisville, non-construction



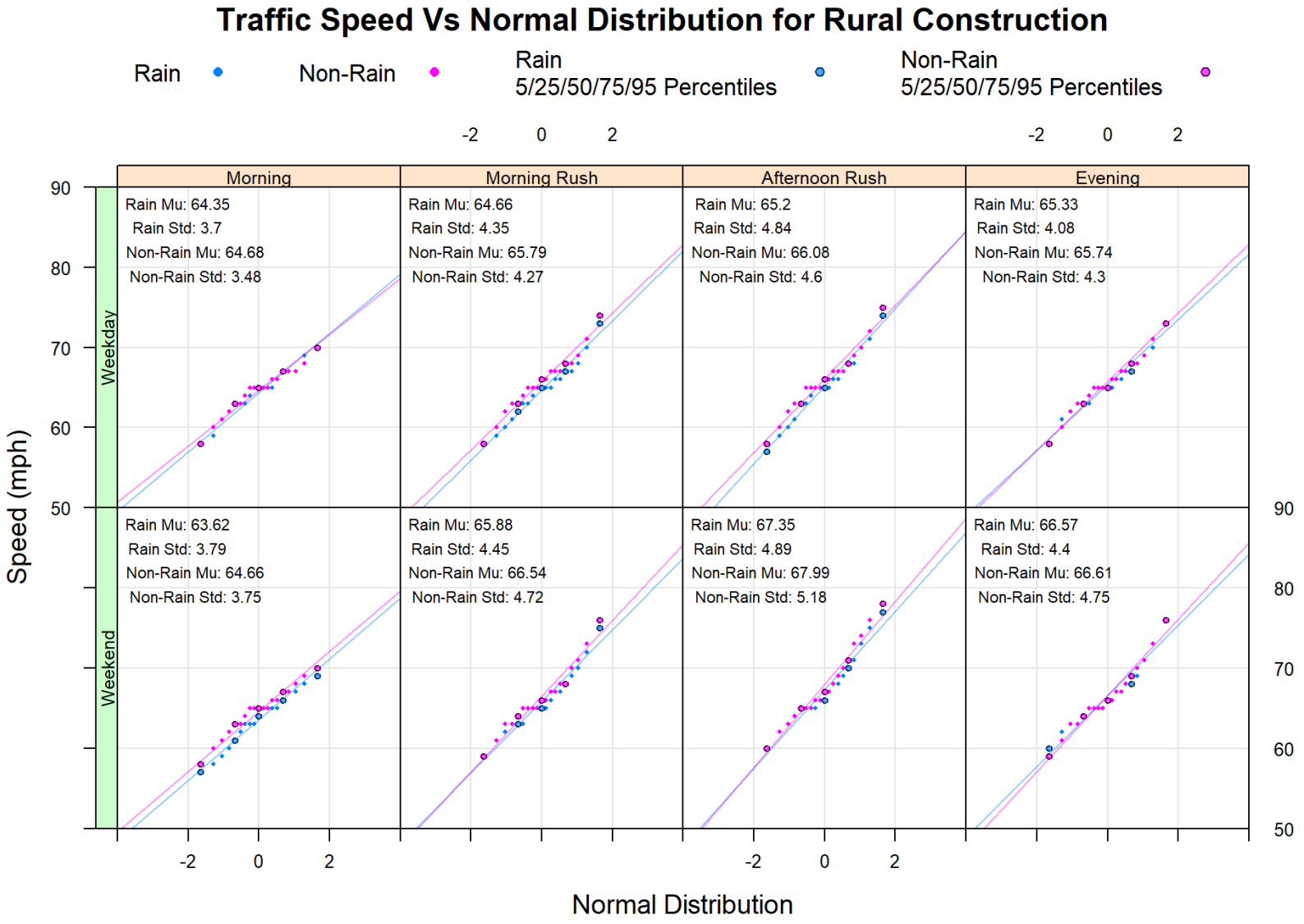
Appendix A 9 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Northern Indiana, construction



Appendix A 10 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Northern Indiana, non-construction



Appendix A 11 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Rural areas, non-construction



Appendix A 12 Normal distribution fits of rain and non-rain traffic speeds above 50 mph for Rural areas, construction

REFERENCES

Center for Advanced Transportation Technology. (2012). *I-95 Corridor Coalition Vehicle Probe Project Guide for Posting Travel Times on Changeable Message Signs Vehicle Probe Project Guide for Posting Travel Times on Changeable Message Signs Report Update*. (June).

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, *13-17-Augu*, 785–794. https://doi.org/10.1145/2939672.2939785

Climate Prediction Center. (n.d.). wgrib2: wgrib for GRIB-2. Retrieved July 31, 2019, from https://www.cpc.ncep.noaa.gov/products/wesley/wgrib2/

Dey, P. P., Chandra, S., & Gangopadhaya, S. (2006). Speed distribution curves under mixed traffic conditions. *Journal of Transportation Engineering*, *132*(6), 475–481. https://doi.org/10.1061/(ASCE)0733-947X(2006)132:6(475)

Federal Highway Administration. (2018). How Do Weather Events Impact Roads? Retrieved July 22, 2019, from https://ops.fhwa.dot.gov/weather/q1_roadimpact.htm

Harmon, T., Bahar, G., & Gross, F. (2018). *FHWA Safety Program Crash Costs for Highway Safety Analysis*. Retrieved from http://safety.fhwa.dot.gov

Ibrahim, Am., & Hall, F. (1994). *Effect of Adverse Weather Conditions on Speed-Flow-Occupancy Relationships*.

Iowa Environmental Mesonet MRMS Grib Archive. (n.d.). Retrieved from https://mtarchive.geol.iastate.edu/2018/06/

Kim, S., & Coifman, B. (2014). Comparing INRIX speed data against concurrent loop detector stations over several months. *Transportation Research Part C: Emerging Technologies*, *49*, 59–72. https://doi.org/10.1016/j.trc.2014.10.002

Ly, S., Degré, A., & Charles, C. (2013). Different methods for spatial interpolation of rainfall data for operational hydrology and hydrological modeling at watershed scale. A review. *Biotechnology, Agronomy and Society and Environment*, *17*(2), 392–406. https://doi.org/10.6084/m9.figshare.1225842.v1

Sathiaraj, D., Punkasem, T.-O., Wang, F., & Seedah, D. P. K. (2018). Data-Driven Analysis on the Effects of Extreme Weather Elements on Traffic Volume in Atlanta, GA, USA. *Computers, Environment and Urban Systems*, *72*, 212–220. https://doi.org/10.1016/j.compenvurbsys.2018.06.012

Shepard, D. (1968). A Two-Dimensional Interpolation Function for Irregularly-Spaced Data. *Proc 23rd Nat Conf*, 517–524. https://doi.org/10.1145/800186.810616

Smith, B. L., Byrne, K. G., Copperman, R. B., Hennessy, S. M., & Goodall, N. J. (2003). An Investigation into the Impact of Rainfall Freeway Traffic Flow. *Review Literature And Arts Of The Americas*.

Stout, G. E. ., & Mueller, E. A. (1968). Survey of Relationships Between Rainfall Rate and Radar Reflectivity in the Measurement of Precipitation. *Journal of Applied Meteorology*, *7*(3), 465–474.

Tanner, J. C. (1952). Effect of Weather on Traffic Flow. *Nature*, Vol. 169, p. 107. https://doi.org/10.1038/169107a0

Tsirigotis, L., Vlahogianni, E. I., & Karlaftis, M. G. (2012). Does Information on Weather Affect the Performance of Short-Term Traffic Forecasting Models? *International Journal of Intelligent Transportation Systems Research*, *10*(1), 1–10. https://doi.org/10.1007/s13177-011-0037-x

U.S. Bureau of Labor Statistics. (n.d.). U.S. Bureau of Labor Statistics: CPI Inflation Calculator. Retrieved December 3, 2020, from https://www.bls.gov/data/inflation_calculator.htm

Zhang, J., Howard, K., Langston, C., Kaney, B., Qi, Y., Tang, L., … Kitzmillller, D. (2016). Multi-Radar Multi-Sensor (MRMS) quantitative precipitation estimation: Initial operating capabilities. *Bulletin of the American Meteorological Society*, *97*(4), 621–638. https://doi.org/10.1175/BAMS-D-14-00174.1