AFRL-RH-WP-TR-2011-0070

USE OF MAHALANOBIS DISTANCE FOR DETECTING OUTLIERS AND OUTLIER CLUSTERS IN MARKEDLY NON-NORMAL DATA: A VEHICULAR TRAFFIC EXAMPLE

Rik Warren
Anticipate & Influence Behavior Division
Behavior Modeling Branch

Robert E. Smith
SRA International
5000 Springfield Street
Dayton OH 45431

Anne K. Cybenko
University of Dayton Research Institute
300 College Park
Dayton OH 45469

JUNE 2011
Interim Report

AIR FORCE RESEARCH LABORATORY
711TH HUMAN PERFORMANCE WING,
HUMAN EFFECTIVENESS DIRECTORATE,
WRIGHT-PATTERSON AIR FORCE BASE, OH 45433
AIR FORCE MATERIEL COMMAND
UNITED STATES AIR FORCE

NOTICE AND SIGNATURE PAGE

Using Government drawings, specifications, or other data included in this document for any purpose other than Government procurement does not in any way obligate the U.S. Government. The fact that the Government formulated or supplied the drawings, specifications, or other data does not license the holder or any other person or corporation; or convey any rights or permission to manufacture, use, or sell any patented invention that may relate to them.

This report was cleared for public release by the 88th Air Base Wing Public Affairs Office and is available to the general public, including foreign nationals. Copies may be obtained from the Defense Technical Information Center (DTIC) (http://www.dtic.mil).

AFRL-RH-WP-TR-2011-0070 HAS BEEN REVIEWED AND IS APPROVED FOR PUBLICATION IN ACCORDANCE WITH ASSIGNED DISTRIBUTION STATEMENT.

RICHARD WARREN
Work Unit Manager
Behavior Modeling Branch

DAVID G. HAGSTROM
Anticipate & Influence Behavior Division
Human Effectiveness Directorate
711th Human Performance Wing
Air Force Research Laboratory

This report is published in the interest of scientific and technical information exchange, and its publication does not constitute the Government's approval or disapproval of its ideas or findings.

REPORT DOCUMENTATION PAGE			*Form Approved* OMB No. 0704-0188
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing this collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to any penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number. **PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.**			
1. REPORT DATE *(DD-MM-YYYY)* 01-06-2011	**2. REPORT TYPE** Interim		**3. DATES COVERED** *(From - To)* March 2009 – June 2011
4. TITLE AND SUBTITLE Use of Mahalanobis Distance for Detecting Outliers and Outlier Clusters in Markedly Non-Normal Data: A Vehicular Traffic Example			**5a. CONTRACT NUMBER** FA8650-09-D-6939 TO0023
			5b. GRANT NUMBER
			5c. PROGRAM ELEMENT NUMBER 62202F
6. AUTHOR(S) [1]Rik Warren, [2]Robert E. Smith, [3]Anne K. Cybenko			**5d. PROJECT NUMBER** 7184
			5e. TASK NUMBER X2
			5f. WORK UNIT NUMBER 7184X21W
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) [2]SRA International [3]University of Dayton Research Institute 5000 Springfield Street 300 College Park Dayton OH 45431 Dayton OH 45469			**8. PERFORMING ORGANIZATION REPORT NUMBER**
9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES) [1]Air Force Materiel Command Air Force Research Laboratory 711th Human Performance Wing Human Effectiveness Directorate Anticipate & Influence Behavior Division Behavior Modeling Branch Wright-Patterson AFB OH 45433-7022			**10. SPONSOR/MONITOR'S ACRONYM(S)** 711 HPW/RHXB
			11. SPONSOR/MONITOR'S REPORT NUMBER(S) AFRL-RH-WP-TR-2011-0070
12. DISTRIBUTION / AVAILABILITY STATEMENT Distribution A: Approved for public release; distribution is unlimited.			
13. SUPPLEMENTARY NOTES 88ABW/PA cleared on 27 Jun 2011, 88ABW-2011-3663.			

14. ABSTRACT

Modeling the behavior of interacting humans in routine but complex activities has many challenges, not the least of which is that humans can be both purposive and negligent, and further can encounter unexpected environmental hazards requiring fast action. The challenge is to characterize and model the humdrum routine while at the same time capturing the deviations and anomalies which arise from time to time. Because of the disruptive impact that anomalies (such as accidents) can have and the importance for incorporating them in our models, this report focuses on one technique for identifying anomalies in complex behavior patterns especially when there is no sharp demarcation between routine and unusual activity. The technique we evaluate is that of Mahalanobis distance which is known to be useful for identifying outliers when data is multivariate normal. But, the data we use for evaluation is deliberately markedly non-multivariate normal since that is what we confront in complex human systems. Specifically, we use one year's (2008) hourly traffic-volume data on a major multi-lane road (I-95) in one location in a major city (New York) with a dense population and several alternate routes. The traffic data is rich, large, incomplete, and reflects the effects of bad weather, accidents, routine fluctuations (rush hours versus dead of night), and one-time social events. The results show that Mahalanobis distance is a useful technique for identifying both single-hour outliers and contiguous-time clusters whose component members are not, in themselves, highly deviant.

15. SUBJECT TERMS

Mahalanobis Distance, Outlier Detection, Outlier Cluster Detection, Vehicular Traffic Analysis, Non-Normal Multivariate Data Analysis

16. SECURITY CLASSIFICATION OF: UNCLASSIFED			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON** Richard Warren
c. REPORT U	**b. ABSTRACT** U	**c. THIS PAGE** U	SAR	60	**19b. TELEPHONE NUMBER** *(include area code)* NA

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. 239.18

THIS PAGE LEFT INTENTIONALLY BLANK

TABLE OF CONTENTS

List of Figures

List of Tables

1 SUMMARY

Modeling the behavior of interacting humans in routine but complex activities has many challenges, not the least of which is that humans can be both purposive and negligent, and further can encounter unexpected environmental hazards requiring fast action. The challenge is to characterize and model the humdrum routine while at the same time capturing the deviations and anomalies which arise from time to time. Because of the disruptive impact that anomalies (such as accidents) can have and the importance for incorporating them in our models, this report focuses on one technique for identifying anomalies in complex behavior patterns especially when there is no sharp demarcation between routine and unusual activity.

The technique we evaluate is that of Mahalanobis distance which is known to be useful for identifying outliers when data is multivariate normal. But, the data we use for evaluation is deliberately markedly non-multivariate normal since that is what we confront in complex human systems. Specifically, we use one year's (2008) hourly traffic-volume data on a major multi-lane road (I–95) in one location in a major city (New York) with a dense population and several alternate routes. The traffic data is rich, large, incomplete, and reflects the effects of bad weather, accidents, routine fluctuations (rush hours versus dead of night), and one-time social events.

The results show that Mahalanobis distance is a useful technique for identifying both single-hour outliers and contiguous-time clusters whose component members are not, in themselves, highly deviant.

2 INTRODUCTION

People walking, driving cars, and flying airplanes are interacting entities whose behavior is typically purposeful and generally characterized by an avoidance of collisions.

Such dynamically interacting agents usually comply reasonably well with "soft" constraints such as traffic speed limits, one-way streets, and parking regulations. Soft constraints can also include unstated social "conventions" such as pedestrians on a sidewalk usually walking on the right of on-coming people, and cars discouraged from passing on the right of a car traveling in the same direction.

"Usually comply" does not mean "always comply" and this means that accidents sometimes happen. Further, accidents can happen due to factors beyond a driver's control such as environmental surprises including a heavy rain, ice, a large pot-hole, or a large animal darting across a road.

One conclusion is that the collective behavior of relatively-autonomous softly-constrained independent *purposeful* agents in real environments will be hard to understand. But we do want to understand how people behave in routine and non-routine circumstances. That means that in addition to the mundane, our models must capture the types, effects, and frequencies of anomalies in everyday life. Toward that end, this report focuses on exploring one technique for identifying anomalies—especially when there is no sharp demarcation between routine and unusual activity—and their frequencies in one facet of daily life, namely, highway traffic. But although traffic analysis is our exemplar, our aim is the more general problem of modeling complex purposive behavior and anomalies.

2.1 Modeling Complex Purposive Behavior & Anomalies

Modeling the behavior of interacting humans in routine but complex activities has many challenges, not the least of which is that humans can be both purposive and negligent, and further can encounter unexpected environmental hazards requiring fast action. The challenge is to characterize and model the humdrum routine while at the same time capturing the deviations and anomalies which arise from time to time.

The humdrum and routine collective behavior of ants, termites, and other insects can create complex structures, order, and patterns which, as research in the area of self-organization in biological systems shows (Camazine, Franks, Sneyd, Bonabeau, & Deneubourg, 2003), can be effectively modeled. But the key term here is "purposeful."

Purposefulness, together with its evil relative "negligence," can be argued to be disruptive of natural order. Even well-motivated well-intentioned persons exercising carefully planned and direct control in a complex situation can unleash a rare but devastating catastrophe. Such anomalies must be assessed for frequency and magnitude before they can be incorporated for in our models. That is, our models of complex interactive human behavior must be able to account for both the routine and the non-routine.

The collective behavior of individuals is not without order and pattern (e.g., Miller & Page, 2007). Even in unregulated dense merging street traffic, there can be order albeit in apparent chaos (Vanderbilt, 2008). Within regulated highways, there are fascinating emergent wavefronts and the models and simulations of them are quite striking (e.g., Beaty, 2011).

Developing a model for well-ordered behavior is one thing, but developing a model which can account for disruptions requires much more than model sophistication. We need to have data on the frequency and types of anomalies and their effects. Hence, as a prerequisite for model building, this report focuses on one technique for identifying anomalies in complex behavior patterns especially when there is no sharp demarcation between routine and unusual activity.

2.2 Identifying Individual Outliers & Anomalies

The technique we evaluate is that of Mahalanobis distance (Mahalanobis, 1936) which is known to be useful for identifying outliers when data is multivariate normal. But, the data we use for evaluation is deliberately markedly non-multivariate normal since that is what we confront in complex human systems.

Since the problem of *outlier detection* is well-studied and well-discussed in statistics and data mining (e.g., Barnett & Lewis, 1994; Johnson & Wichern, 2007; Shekhar, Lu & Zhang, 2003; Rousseeuw & Leroy, 1987), what is unique or interesting about the current report and analysis? Our concern is with:

- Identifying outlier contiguous *clusters* and not just stopping with identifying *individual* outliers;

- Determining usefulness in spite of assumption failures; and

- Analyzing traffic-volume data to illustrate the techniques.

2.3 Identifying Anomalous Contiguous Clusters

Typically, the unit of analysis or "unit of data" or data "record" is generally fixed or given, and the typical task focuses on identifying unusual records. More graphically, the usual problem is to identify individual outlier *points* in a (multi-dimensional) scatter plot. In the cases of interest in this report, the (only) available data-records are indeed traditional data-"points," but the emphasis is on identifying spatially or temporarily contiguous *clusters* of points or data-records. It may well be that none of the members of an outlier cluster are particularly deviant as individual points.

Note that the emphasis is on multiple data-records in the discussion of clusters. This should not be confused with extremeness on the variables *within* a data-record. As Barnett and Lewis (1994, p. 270) point out: "A multivariate outlier need not be extreme in any of its components. Someone who is short and fat need not be the shortest, or the fattest, person around." For an illustration, see Figure 2.

2.4 Robustness Considerations

Another feature of this report is its concern with determining usefulness in spite of assumption failures. Instead of dwelling on probabilities based on theoretical distributions, We advocate the use of percentiles and comparison with an external information which can provide "ground truth." Since no one method solves all problems, we need to know when a method such as Mahalanobis distance breaks down and its use contraindicated.

2.5 Traffic Volume Analysis

Traffic data, in general, is especially well-suited to study the effects of purposive and negligent behavior in complex situations due to the variety of choices and actions which can be, and often must be, made. Purposeful and somewhat unconstrained traffic behavior by sentient drivers includes conscious choice of roads ("surface streets") or highways, direction of travel, speed (including too fast or too slow), lane (including choosing to drive slowly in a fast lane), lane changes, highway entrance and exit, and time of travel. Yet another interesting voluntary behavior is *rubbernecking*—slowing down to watch an accident scene.

Traffic volume data, in particular, is especially interesting to illustrate the problem of cluster outlier identification. The data records used here are hourly traffic volumes crossing a particularly active location on a major Interstate Highway. Yet, the events of interest—such as accidents—often span several hours, hence we are faced with the higher-order problem of identifying unusual patterns beyond single outliers. Also, the data-set is real and hence generally not (multivariate-)normal. Further, the data-set has missing records which always presents challenges. The analysis challenge is increased since ancillary data, such as weather and accident records, are not always available to establish "ground truth."

3 MAHALANOBIS DISTANCE

Given a theoretical distribution or an empirical data set and a particular data point, a basic question concerns the "rarity" or extremeness of the datapoint relative to the other data points. How distant is the point from the center of the distribution? How likely or unlikely is it that the point lies at that particular distance (or closer) from the center of the distribution?

For univariate data, familiar Euclidean distance and rarity are easily related, but for multivariate data, the concept of distance, as we will see, must be modified in order to relate "distance" and rarity. That requisite modification is Mahalanobis distance. It enables a powerful technique for detecting multivariate outliers. But before presenting this general multivariate method, we briefly review some basic methods for identifying univariate and bivariate outliers and anomalous data points.

3.1 Detecting Univariate Outliers & Rarity

In elementary statistics, we learned to transform raw scores into z-scores using the mean and standard deviation of the data by:

$$z = \frac{\text{score} - \text{mean}}{\text{standard deviation}} = \frac{x - \mu}{\sigma} \tag{1}$$

The z-scores are the *signed* distances of the datapoints from the mean in units of the standard deviation. They simply rescale the data so that the mean is zero and the standard deviation is one. Hence, z-scores preserve the shape of the original distribution. Furthermore, since the units of the numerator and denominator are the same, z-scores are dimensionless.

If the distribution is normal, the probability of observing a particular z-score (or one closer to the mean of zero) is easily calculated or obtained from a table of so-called "normal deviates." Thus, distance and rarity are solidly linked.

Even if the distribution is *not* normal, **Tchebycheff's** (a.k.a. Chebyshev's) **Inequality** assures us that there is still a reasonable linkage between rarity and the (unsigned) distance of a point from the mean:

- The probability of a z-score with an absolute value greater than or equal to, say, k, must be less than or equal to $1/k^2$. For example, no matter what the distribution of the data, the probability of observing a z-score of 5 or greater is no greater than $1/5^2$ or .04. And the closer the distribution is to normal, the probability could be considerably less. Hence, a z-score of 5 is always unusual enough to consider as a possible outlier.

The following transform of Equation 1 is not in elementary books, but enables a bridge or generalization to multivariate data. First, square the z-score:

$$z^2 = \frac{(x - \mu)^2}{\sigma^2}$$

Next, rewrite the righthand side to produce

$$z^2 = (x - \mu)(\sigma^2)^{-1}(x - \mu) \tag{2}$$

Admittedly, Equation 2 provides no computational advantage, but its peculiar form is the form necessary for generalization to multivariate data.

3.2   Detecting Bivariate Outliers & Rarity

A bivariate datapoint or observation consists of two linked values, x and y, drawn from two populations, X and Y. The pair of values may be represented by a column vector

$$\vec{p} = \mathbf{p} = \begin{bmatrix} x \\ y \end{bmatrix} = [x, y]'$$

where the prime signifies *transpose*[1] in statistics.

The means of the X and Y distributions, taken together, define the *centroid*:

$$\vec{\mu} = \boldsymbol{\mu} = \begin{bmatrix} \mu_x \\ \mu_y \end{bmatrix} = [\mu_x, \mu_y]' \tag{3}$$

Now assume the two independent populations, X and Y, representing two independent variables, are each normally distributed but with unequal standard deviations σ_x and σ_y. Further assume that the variables are uncorrelated so that the covariance σ_{xy} is zero.

We now examine two types of distance of a point $\vec{p} = [x, y]'$ to the centroid $\vec{\mu} = [\mu_x, \mu_y]'$. One is Euclidean and the other is a non-Euclidean distance useful in statistics.

3.2.1   Euclidean Distance to Centroid

If the means of both the X and Y distributions are both zero, the Euclidean distance from the point $[x, y]'$ to the centroid located at the origin is the familiar

$$d = \sqrt{x^2 + y^2}$$

but since, in general, the means of the variables will not be zero, the centroid will not be at the origin but rather at

$$[\mu_x, \mu_y]'$$

so the Euclidean distance of a point to the centroid $\vec{\mu} = [\mu_x, \mu_y]'$ is:

$$d = \sqrt{(x - \mu_x)^2 + (y - \mu_y)^2} \tag{4}$$

Equation 4 is technically correct but not completely satisfactory from a statistics point of view.

- If the two variables carry different units, e.g., one in miles per hour and one in fatalities, it is not clear what the units for the two-dimensional distance should be.

- Even if the two variables carry the same units, Equation 4 is sensitive to the scale used so the distance values can vary wildly with different scale units. This complicates evaluating what constitutes an extreme value.

---

[1]Note that $[x, y]'$ is a *column* vector despite appearances. A prime on a column-vector transposes it into a *row vector*:

$$\vec{p}' = \mathbf{p}' = \begin{bmatrix} x \\ y \end{bmatrix}' = [x, y]$$

The transpose of column vectors is common in statistics to conform to the rules of vector and matrix operations.

- Equation 4 ignores information provided by the standard deviations. Even if the variables have the same units, standard deviations can reflect the quality of the measurement process. If the trustworthiness of the two variables is not the same, it is inappropriate to give them the same weight which is what Equation 4 does.

- The locus of all pairs of points $[x, y]$ at the same distance from the centroid is a circle since setting d to a constant yields the equation of such a circle. This is a graphic realization that Equation 4 gives equal weights to the two variables.

3.2.2 Non-Euclidean Weighted Distance to Centroid

The inadequacies of Equation 4 can be finessed by appealing to the logic behind the use of standard scores (see Eq. 1). Instead of using $(x - \mu_x)$ and $(y - \mu_y)$ as components in a distance equation, first weigh each difference by the inverse of its associated standard deviation. That is, compute weighted distances using z-scores instead of the differences from the means of the variables:

$$d_w = \sqrt{z_x^2 + z_y^2} \tag{5}$$

or, equivalently:

$$d_w = \sqrt{\frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(x - \mu_y)^2}{\sigma_y^2}} \tag{6}$$

This procedure counters the objections raised about the unweighted Euclidean distance:

- Since z-scores are are dimensionless, the problem of mixing apples and oranges disappears.

- Since z-scores rescale everything to standard deviation units, the problem of different scales per variable disappears and assessment of extremeness can be based on a function of the z-scores, namely, χ^2. (See Section 3.4.)

- Rescaling scores by the inverse of the standard deviations means that the more variable factors are given less weight thereby incorporating information provided by the magnitude of the standard deviation.

To better appreciate the weighted distance, it is useful determine the locus of all points having the same weighted distance from the centroid. Begin by squaring Equation 6:

$$d_w^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(x - \mu_y)^2}{\sigma_y^2}$$

and then dividing both sides by d_w^2 to get:

$$1 = \frac{(x - \mu_x)^2}{\sigma_x^2 d_w^2} + \frac{(x - \mu_y)^2}{\sigma_y^2 d_w^2} \tag{7}$$

Equation 7 is just the equation of an ellipse centered at the centroid $\vec{\mu} = [\mu_x, \mu_y]'$ and having axes of length $\sigma_x d_w$ and $\sigma_y d_w$. Since Equation 7 is equivalent to Equation 5, the iso-weighted-distance contours defined by Equation 7 tell us that points located along the same ellipse can

7

be considered to have the same statistical "closeness" or "extremeness" from the centroid even though their Euclidean distances can differ. Indeed, the ellipses can be interpreted as probability density contours. (See Section 3.4.)

3.2.3 Alternate Computational Forms

As was the case with univariate data, it is useful to work with squared distances and to seek alternate computational forms. Squaring Equations 5 and 6 yields:

$$d_w^2 = z_x^2 + z_y^2 \tag{8}$$

or, equivalently:

$$d_w^2 = \frac{(x - \mu_x)^2}{\sigma_x^2} + \frac{(x - \mu_y)^2}{\sigma_y^2} \tag{9}$$

Equation 9 can be "vectorized" as follows:

- Form the column vector

$$[\vec{p} - \vec{\mu}] = \left[\begin{array}{c} x - \mu_x \\ y - \mu_y \end{array} \right]$$

- Form the covariance matrix

$$\boldsymbol{\Sigma} = \left[\begin{array}{cc} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{array} \right]$$

- Form the inverse of the covariance matrix $\boldsymbol{\Sigma}$. Since, in this special case, $\boldsymbol{\Sigma}$ is diagonal:

$$\boldsymbol{\Sigma}^{-1} = \left[\begin{array}{cc} 1/\sigma_x^2 & 0 \\ 0 & 1/\sigma_y^2 \end{array} \right]$$

- Lastly, set

$$d_w^2 = [(x - \mu_x), (y - \mu_y)] \left[\begin{array}{cc} 1/\sigma_x^2 & 0 \\ 0 & 1/\sigma_y^2 \end{array} \right] \left[\begin{array}{c} x - \mu_x \\ y - \mu_y \end{array} \right] \tag{10}$$

When the vector and matrix multiplications in Equation 10 are carried out, the result is Equation 9. Note that Equation 10, and thereby, Equation 9, can be written compactly as:

$$d_w^2 = [\vec{p} - \vec{\mu}]' \, \boldsymbol{\Sigma}^{-1} [\vec{p} - \vec{\mu}] \tag{11}$$

which has structural and content affinities with Equation 2.

3.2.4 Generalizing Beyond Two (Uncorrelated) Variables

The form of Equations 8, 9 and 10 reveals a pattern that may be used to include a third variable, z, and more *assuming all variables are uncorrelated*. But for adding variables beyond z, the current notation quickly becomes unwieldly. Two enhancements to enable generalization beyond two variables and a more compact notation are:

- First, replace x, y, and presumably z and beyond, with x_1, x_2, x_3, etc., and z_x, z_y, with z_1, z_2, etc. From here on this is how variables will be denoted.

- Second, capitalize on the new notation for variables and replace the string of similar-patterned terms using an index and a summations sign.

Thus:

$$d_w^2 = \sum z_{x_i}^2 \tag{12}$$

or, equivalently:

$$d_w^2 = \sum \frac{(x_i - \mu_{x_i})^2}{\sigma_{x_i}^2} \tag{13}$$

If all the covariances among the variables are zero, then by the same procedure used to derive Equation 7, Equations 12 and 13 can be put in the standard form of ellipsoids and hyper-ellipsoids depending on the number of variables. And, as was the case with Equation 7, if d_w is set to a constant, the resulting (hyper-)ellipsoid is an iso-weighted-distance shell and can be interpreted as a probability density shell.

Since Equation 12 is a sum of z-scores, one for each variable, it is also the equation of a χ^2 probability distribution which becomes the key to evaluating the rarity of the weighted-distances.

By the same procedure leading to rewriting Equation 9 as Equation 11, Equation 13 can also be rewritten as Equation 9 with no change in notation needed! The vectors and matrices get larger as the number of variables grows, but the basic pattern is the same.

So far, all variables have been assumed uncorrelated and all covariances assumed zero. This is highly unrealistic with real data. This limitation is eliminated by the distance metric developed by Mahalanobis (1936). Further, his distance metric applies to the weighted distance between *any* two multivariate points, not just between a point and the centroid.

3.3 Concept & Definition of Mahalanobis Distance

The basic idea behind Mahalanobis distance is shown in Figure 1. Assume the ellipse represents the outline of a bivariate-normal scatterplot. Although all points on the circle are equidistant from the center, a point on the circle on the short axis is statistically more "deviant" (in terms of its probability of being that far-out in the periphery of the scatterplot) than a point on the circle along the long axis (which being close to the center of the scatterplot relative to the long-axis has a high-probability of occurrence). Statistically speaking, all the points on an ellipse are argued to be equiprobable and thus, statistically equidistant from the center of the ellipse. In fact, some (such as Johnson & Wichern, 2007) just use the term "statistical distance" or "standard distance" (e.g., Flury, 1997) or even "elliptical distance." A set of nested concentric ellipses, each corresponding to one probability value, are referred to as *probability density ellipses*.

In general, if $\vec{x} = [x_1, x_2, \ldots, x_p]^T$ and $\vec{y} = [y_1, y_2, \ldots, y_p]^T$ are multivariate data-points (or observations or records or cases) drawn from a set of p variables with a $p \times p$ covariance matrix $\mathbf{S}$, then the Mahalanobis distance d_m between them is defined as:

$$d_m(\vec{x} - \vec{y}) = \sqrt{(\vec{x} - \vec{y})^T \mathbf{S}^{-1} (\vec{x} - \vec{y})} \tag{14}$$

In spite of the straight-forward definition, a few points need to be kept in mind:

Euclidian vs. Mahalanobis Distance
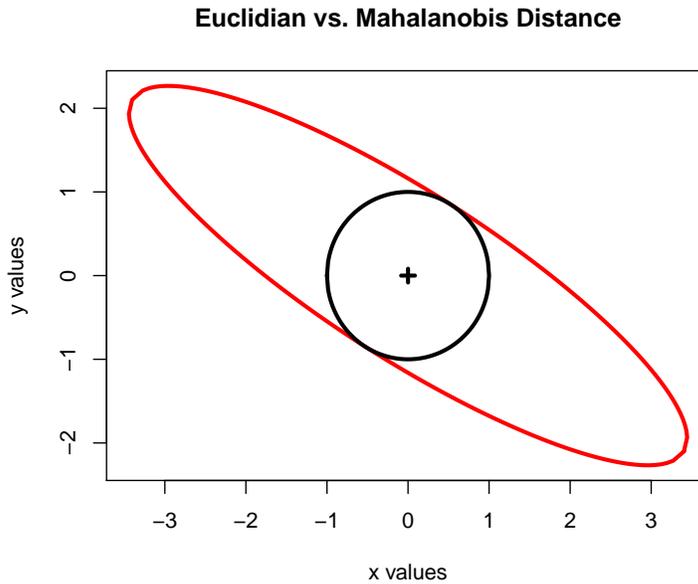
Figure 1: Euclidian vs. Mahalanobis Distance

- Although, the motivation behind the definition appeals to ellipses and multivariate normal distributions, the definition is silent about the underlying p distributions from which the data vectors are drawn.

- There is no standard notation for Mahalanobis distance in the literature.

- Many authors prefer to work with the *square* of the Mahalanobis distance. There are some benefits to the use of this so-called "generalized distance":

  - The resulting equation, if written out, has the appearance of the canonical form of the equation for an ellipse.

  - The squared value, for multivariate-normal data, is intimately related to the χ^2 distribution.

- One drawback is that a few authors appear to erroneously refer to the squared-value as the Mahalanobis distance.

The usefulness of Equation 14 depends on the distributions of the p variables and also the specifics of the reference or comparison vector $\vec{y}$.

3.4 Multivariate Normal Distributions & Ellipses

In particular, if the underlying distribution of the p random variables is exactly multivariate normal with with a $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ and if $\vec{y} = [y_1, y_2, \ldots, y_p]^T$ has the constant value $\vec{\mu} = [\mu_1, \mu_2, \ldots, \mu_p]^T$ formed from the population means of the p random variables then the Mahalanobis distance d_m of a particular multivariate data-point $\vec{x}$ from $\vec{\mu}$ is:

$$d_m(\vec{x}) = \sqrt{(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu})} \tag{15}$$

Setting d_m to a constant c defines a multidimensional ellipsoid with centroid at $\vec{\mu}$. The shell of the ellipsoid is a probability density contour, and the probability associated with each c^2 (not c) shell follows a χ^2 distribution with p degrees of freedom. As Johnson and Wichern (2007, p. 155, Eq. 4-8) state: "The *solid* ellipsoid of $\vec{x}$ values satisfying

$$(\vec{x} - \vec{\mu})^T \Sigma^{-1} (\vec{x} - \vec{\mu}) \leq \chi_p^2(\alpha) \tag{16}$$

has probability $1 - \alpha$." (*emphasis added and some notation changed*)

3.5 Strengths of d_m for Outlier Detection: Usage Notes

Mahalanobis distance is often used to determine multivariate outliers due to its several strengths:

- It features numerical and graphical thresholds to determine outliers.

- It is flexible and permits the use of robust and independent choices for the centroid and covariance matrix values.

- It is amenable to techniques for mitigating the influence of outliers during the search for the outliers.

- It can detect unusual patterns within a multivariate observation.

- It provides an alternative to regression techniques when there is no obvious value to be predicted.

Thresholds for outliers. The theory embodied by Equation 15 is compelling and the relationship to χ^2 facilitates evaluation of "candidate" outliers. For example, for two variables ($p = 2$), and $\alpha = .05$, $\chi^2 = 5.99$. Equation 16 then means that only 5% of squared Mahalanobis distances are expected to be greater than 5.99. Graphically, only 5% of points should lie outside of an ellipse whose contour is defined by setting:

$$d_m = c = \sqrt{c^2} = \sqrt{\chi^2} = \sqrt{5.99} = 2.45$$

Thus, the square root of the critical value of χ^2 can serve as a threshold for determining outliers.

Fexibility. Another reason Mahalanobis distance is useful for outlier detection is its flexibility. The defining equation (Eq. 14) deliberately used $\mathbf{S}$ (an empirical or sample quantity) for the covariance matrix rather than Σ (a theoretical or population quantity). Obviously, in empirical work, a data-based estimator of the centroid must be used, but instead of using the means of the p variables, a researcher can readily use a centroid based on medians or trimmed means.

Mitigating outlier effects while searching for outliers. The problem of determining outliers from a dataset containing outliers is well illustrated by Rousseeuw and Leroy (1987).

Within the context of classical multiple regression, they point out that outliers cannot always be discovered by looking at least squares residuals: Some outliers are leverage points which greatly distort the regression lines or surfaces so that they (the leverage-point outliers) yield relatively small least squares and some "good" points yield large least squares. Roousseuw and Leroy (p. 8) state that "...there are many multivariate data sets ...where the outliers remain invisible even through a careful analysis of the LS residuals." (LS means least squares.)

Within the context of Mahalanobis distance (where least squares is not germane), there is a similar problem of the distortion effects of outliers on the basic technique.

Using medians or trimmed means is one way of mitigating the effects of outliers on the referent data-set. The problem here is that the means-based centroid and the covariance matrix of a data-set are themselves distorted by outliers. But using a, say, median-based centroid does not solve the problem of a distorted covariance matrix.

To mitigate the effects of outliers on both the centroid and the covariance matrix, one technique is to remove the point (data vector) whose Mahalanobis distance is being determined and compute the centroid and covariance matrix from the remaining $n - 1$ points of the n-point full data set. This procedure would be done for each of the n points in turn. If the dataset is large, this procedure requires computing a centroid and covariance matrix—and its inverse—for each of the n data vectors. But if the dataset is very large, then the effect of an outlier on a large mass could be negligible and the extra computational load avoided.

Detecting an Observation Having an Unusual Pattern. Since a multivariate record or observation is a set of numbers, the numbers can form patterns, some of which are typical and some of which are less typical. For example, in a three-variable set, the first two numbers might tend to be high and nearly equal but the third number might tend to be very low. If this is the typical pattern, then an observation in which all three number were roughly equal might be anomalous.

The numbers can be plotted in various ways which make the patterns stand out visually. For example, Figure 17 shows one way to plot a set of three numbers, but it would have been possible to plot the linked numbers as vertical bars or star charts. The linked lines were chosen to emphasize that a typical pattern in that dataset has all three numbers roughly equal but that, at the same time, the relative height of the pattern is not *generally* a factor in typicality.[2]

Use Where Regression Not Obvious. As the above discussion of Mahalanobis distance shows, d_m can be directly used to detect unusual patterns in observations. Regression techniques can also be used to detect outliers (e.g., Rousseeuw & Leroy, 1987). But in the case of certain types of data, such as traffic-volume per lane of a multi-lane highway, what is to be the dependent variable or variables? It is possible to create some variable (e.g., total highway volume or total volume in one direction) to be a function or composite of the independent variables (e.g., the individual lane volumes), but there has to be a cogent reason for the new composite variable and a compelling argument made for why it can indicate useful outliers. For the data analyzed in Section 4, it is the case that total same-direction volume

---

[2] "Generally" is used here to signal caution. For the dataset underlying Figure 17, an observation of three numbers all zero or near zero, would indeed be anomalous.

as such, for example, is not directly linked to anomalies since low total volume is typical in late-night hours and high lane-volumes during a rush hour can still show unusual patterns. Perhaps some composite using ratios of lane volumes or lane-volume differences might be devised as a good indicator composite, but it is not obvious what it would be. Mahalanobis distance avoids such difficulties.

3.6 Weakness of d_m with Non-Multivariate Normal Distributions

If the underlying distribution of one or more of the p variables is not normal, the quadratic form in Equation 14 still yields an ellipse when set to a constant. However, the interpretation of the elliptical shells as probability density contours, and the ellipsoid volumes as representing probabilities following a χ_p^2 distribution is no longer justified.

When the deviation from a multivariate normal distribution is slight, it might be argued that the χ_p^2 probabilities are, at least, suggestive and that the Mahalanobis approach is reasonably robust. But, terms like "slight" and "reasonable" are imprecise.

Moreover, some distributions, such as those which are multimodal, are especially problematic. Figure 2 shows a classic illustration of one problem in outlier identification with multivariate data: Namely, the problem that a mid-valued observation among polarized data can be an outlier.
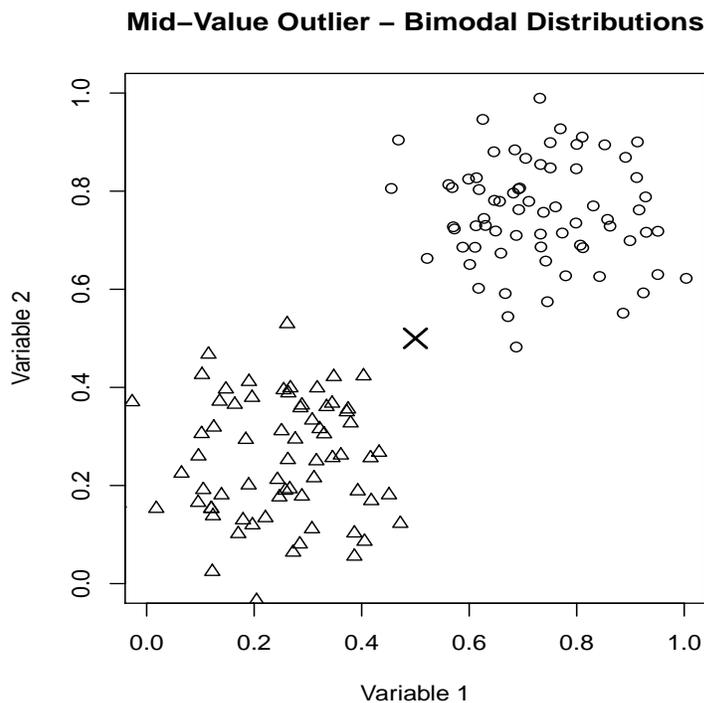

Figure 2: Two Bimodal Variables: Outlier in Middle

As a literally extreme example, if the mid-valued outlier in Figure 2 has the same vector-value as the centroid, i.e., if

$$\vec{x} = \hat{\vec{\mu}} \tag{17}$$

where $\hat{\vec{\mu}}$ is the estimator of $\vec{\mu}$ given by the means of each of the p variables, then the Mahalanobis distance from the centroid is zero. Further discussion of this, and similar, cases is deferred until later.

3.7 Non-Normal Distributions: Statistical Approaches

There are so many ways that individual and compound distributions can deviate from normality that there is no single technique to determine when conclusions based on normality assumptions break down. Several techniques for assessing the assumption of normality are presented in, for example, Barnett and Lewis (1994) and Johnson and Wichern (2007). Johnson and Wichern (p. 177), in particular, note that since "all linear combinations of normal variables are normal and the contours of the multivariate normal density are ellipsoids," it is reasonable to ask if (1) the marginal distributions of the data appear to be normal including the distributions of a few linear combinations of variables, and (2) if the scatter plots of pairs of observations give an elliptical appearance.

Techniques include using histograms, scatter-plot matrices, normal Q-Q plots, correlations coefficient tests, and chi-square plots. There is a predominance of graphical procedures, accompanying visual inspections, and judgment calls by an analyst.

Such procedures are very enlightening, but not readily amenable to automation and thus cumbersome when there are large numbers of cases and variables. The workload is especially compounded when there are many separate analyses to be performed, for example, when analyzing data from many multi-lane traffic sensor locations.

In situations involving large numbers of cases and large numbers of analyses, some deviations from normality are to be expected, and it is reasonable to ask how seriously the deviations should be taken.

3.8 Non-Normal Distributions: Percentile Approach

Let us not lose sight of what we are after: We are after multivariate outliers, but these can, to again use Rousseeuw and Leroy's (1987, p. 8) words, "remain invisible" when there are more than two variables. Our own Figure 2, although bivariate and therefore subject to visual scrutiny, is suggestive and illustrative of the greater-than-two multivariate case. The conclusion is that, in general, visual inspection of plotted data is not always a viable approach to outlier detection.

One non-visual solution is to use a statistical technique and define and outlier as being in the outer 1% or 5% of the assumed distribution function. But, since statistical assumptions are often not met, especially when there are outliers present, the indicated thresholds or cut-off values will not declare exactly 1% or 5% of the data as outliers.

But if visual inspection is barred to us, and we must resort to using percentages to define outliers, why use techniques which can yield actual percentages other than the ones we use to define an outlier?

The approach to outlier definition we advocate here is to declare *exactly* the outer (one- or two-tailed as appropriate) 1% or 5% or some other $x\%$ of the *empirical* values of some summary per-observation index, such a Mahalanobis distance, as outliers. This finesses the problems of assumption failure and robustness. The researcher can then modify the criterion

based on properties of the distribution of the empirical index values. For example, does the histogram of the empirical values show any groupings or gaps indicating a better cut-off for declaring outliers?

4 VEHICULAR TRAFFIC EXAMPLE

Vehicular traffic, whether routine or unusual, is in and of itself an interesting dynamic social system with social and economic consequences beyond just highways. As such, the more we understand traffic flows, the better future highways can be designed for greater safety and fuel economy.

The problem of understanding vehicular traffic patterns on a major highway also affords a nice opportunity to assess the advantages, limitations, and subtleties of using Mahalanobis distances for analyzing social dynamics.

The specific data featured here consists of hourly traffic counts at a particular road location, so the Mahalanobis distance technique will focus on identifying unusual patterns, i.e., outliers, in the hourly traffic counts. But many events of interest, such as accidents or construction disruptions, can span several hours, so techniques to identify outlier clusters which exceed the span of an hour will also be explored.

4.1 Specific Roadway & Traffic Dataset

Traffic data is from a sensor station located on Interstate 95 (I–95) between Jerome Avenue and Webster Avenue in Bronx County in New York City (See Figure 3). The sensor station is 2.66 miles from where I–95 crosses the border from New Jersey on the George Washington Bridge into New York City's Manhattan Island. Between New Jersey and The Bronx, I–95 passes briefly through a relatively non-commercial section of Manhattan far from its commercial heart. This stretch of I–95, also known as the Cross Bronx Expressway, despite its designated North and South lanes, actually runs East and West. It is a major highway, passing through America's largest city and linking two dense population areas northeast and southwest of New York City. Further, the highway and sensor site are surrounded by several significant highways and surface streets. Contrary to expectations for many commercial centers, the site does not experience a pronounced imbalance (although there is some) in traffic in different directions corresponding to a net influx of morning commuters and a net exodus of evening commuters.

The data consists of hourly traffic counts for each of six lanes (three northbound and three southbound) throughout the Year 2008.

4.1.1 Data Source & Source Data

All traffic data were downloaded from the website of the New York State Department of Transportation (New York State DOT, 2011) which specifically states:

> "The data is intended for a user to perform their own analyses."

The NYSDOT website has descriptions of the data collection procedures and the contents and formats of the reference and data files.

- File `NYHeader_easier.csv` contains sensor site information. Information for the I-95 site in The Bronx, which has the unique road segment identifier (RC_ID) of 010003, is found in row 4,592.

- File `vol_2008.csv`, which is 31 MB, contains data from many locations.
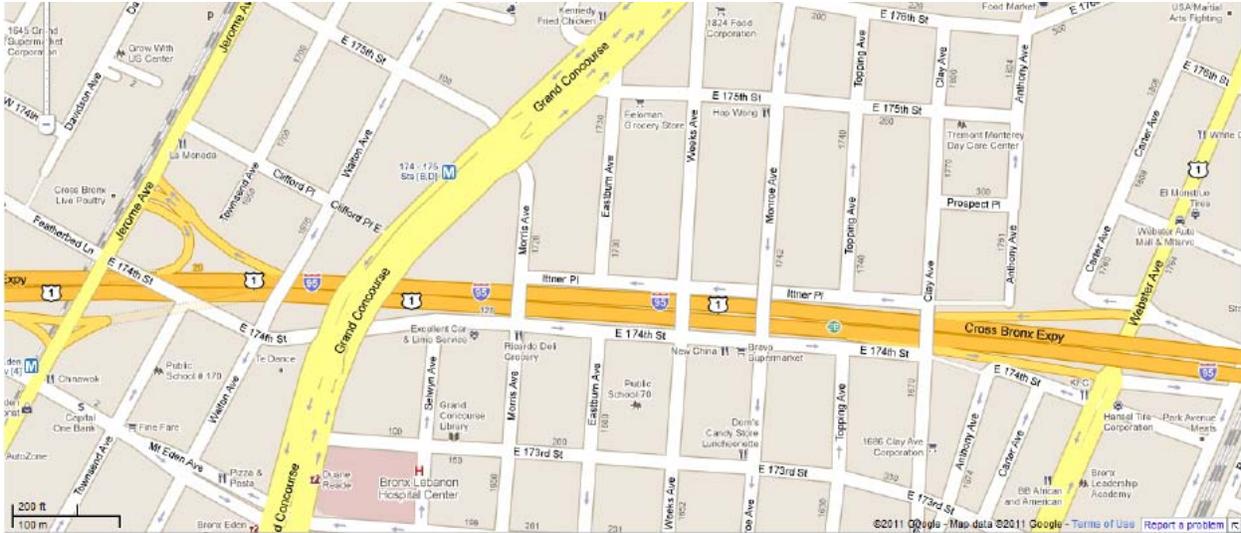
Figure 3: Map of Interstate 95 in The Bronx, New York City

4.1.2 Working Data File Preparation

Since file `vol_2008.csv` contains data from many locations, a new file was created from only those rows containing the value `10003` in the column labeled `RC_ID`. This intermediate file retained the format of the NYSDOT source file which has five columns: RC_ID, Date.Time, Direction, Lane, and Count. This format *generally* results in six rows with the same Date.Time value: one row for each combination of two directions and three lanes per direction. However, sometimes there is missing data in one direction or the other, so sometimes there are only three or no rows for a given Date.Hour. With 366 days and 24 hours per day, there were 8,784 hours in 2008. With 6 rows for each Date.Hour, there should be 52,704 rows of data. The file actually has 48,879 data rows indicating that there are 3,825 traffic counts missing.

After dropping the RC_ID column (since all entries were `10003`), a more "multivariate-analytic"-oriented working data file was then created with each row containing a unique Date.Time value and six columns (labeled S1, S2, S3, N3, N2, N1) containing hourly traffic counts for each of the six traffic lanes or `NA` if volume data was missing for a lane.

4.2 Individual-Lane Traffic Volume Statistics

There were 8,784 hours in 2008 (366 days × 24 hours per day). However, there were 502 hourly traffic volume counts missing per each southbound lane and 773 per each northbound lane. This still leaves large n's per lane for the volume counts: 8,282 southbound and 8,011 northbound.

4.2.1 Basic per-Lane Descriptive Statistics

Table 1 and Figure 4 show the basic traffic hourly volume statistics per lane for the year 2008. For at least one hour, there was no volume on five lanes. Only the middle northbound lane (N2) maintained some traffic on an hourly basis, but the small number of vehicles, 20,

Table 1: Basic Volume Statistics per Lane

	S1	S2	S3	N3	N2	N1
Minimum	0	0	0	0	20	0
Mean	1288	1216	1048	1175	1223	1185
Median	1473	1305	1197	1318	1317	1352
Maximum	2069	1910	1829	2038	1843	1781

suggests that no traffic passed by the sensor for the better part of that low-volume hour. In fact, low volume is rare in both north and south middle lanes as shown by the outliers in Figure 4. There are no outliers on the high end.
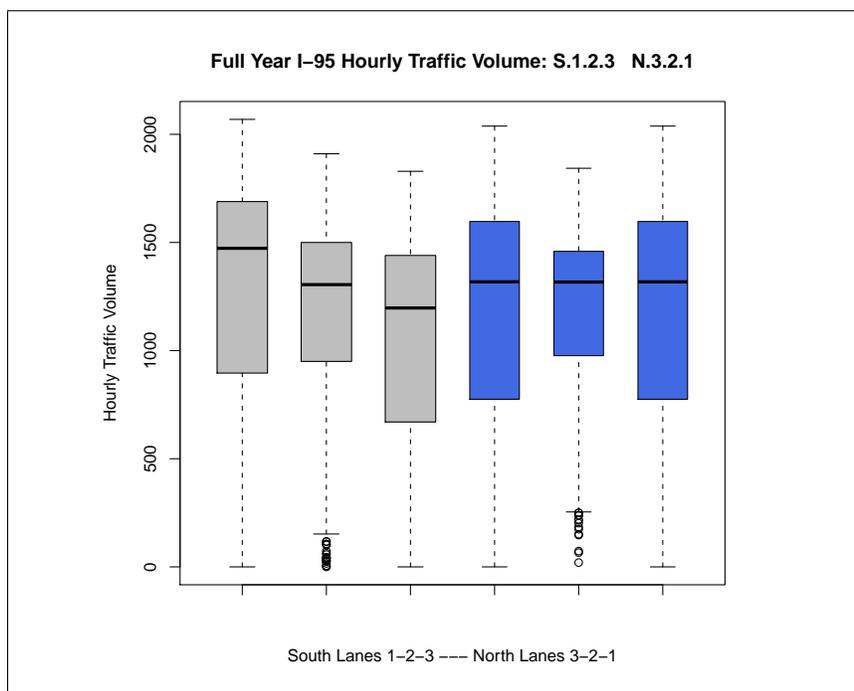
Figure 4: Hourly Traffic: All Six Lanes

4.2.2 Evaluating Normality Assumptions

The relatively high position of the medians on all six boxplots coupled with the relatively shorter upper whiskers suggests that the volume distributions are skewed. And, since one purpose of this analysis is to evaluate the usefulness of Mahalanobis distance under non-normal distributions, it is incumbent to check the empirical distributions more closely. Histograms and normal Quantile-Quantile plots are an effective way to check for normality.

The traffic-volume histograms in Figure 5 clearly show that traffic is asymmetric with high volume more common than low volume. Not only is the distribution asymmetric, but Figure 5 reveals a bi-modality in each histogram. The relative height of the lessor mode,
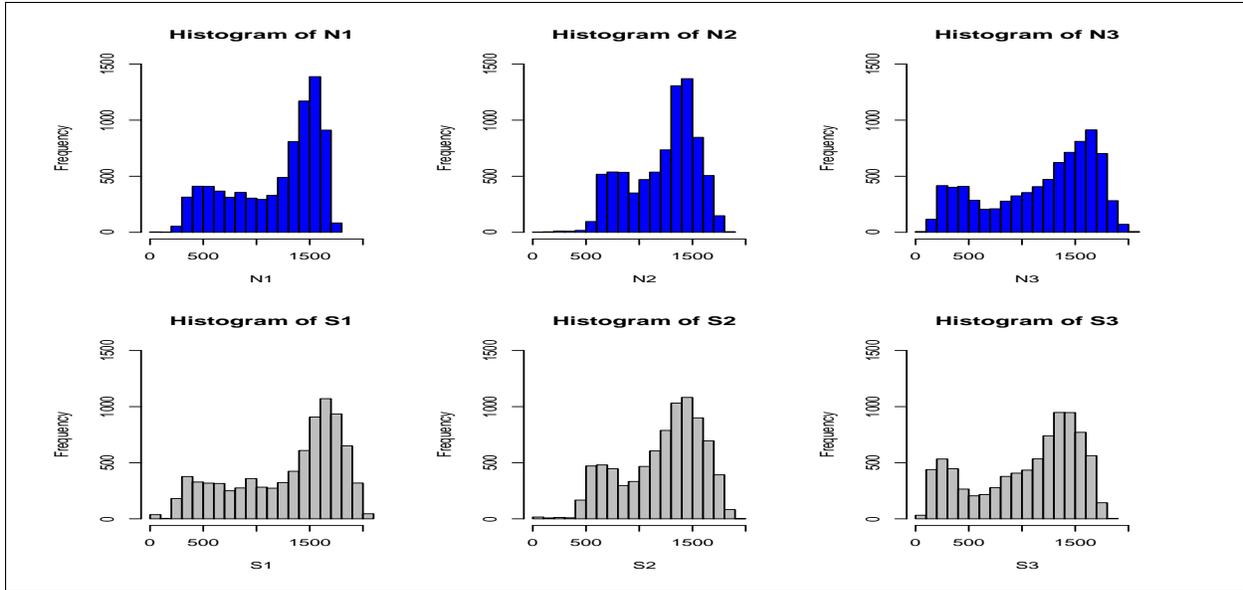
Figure 5: Individual Lane-Volume Histograms

toward the low-volume end, is weak for both (North and South) outer lanes, moderately stronger for both middle lanes, and clearly defined for both left-most lanes.

The conclusion of non-normality due to asymmetric bimodal distributions shown in the histograms is further supported by the markedly non-linear Q-Q plots shown in Figure 6. This means that the χ^2 probabilities and ellipses associated with the (squared) Mahalanobis distance values discussed below are, at best, suggestive.
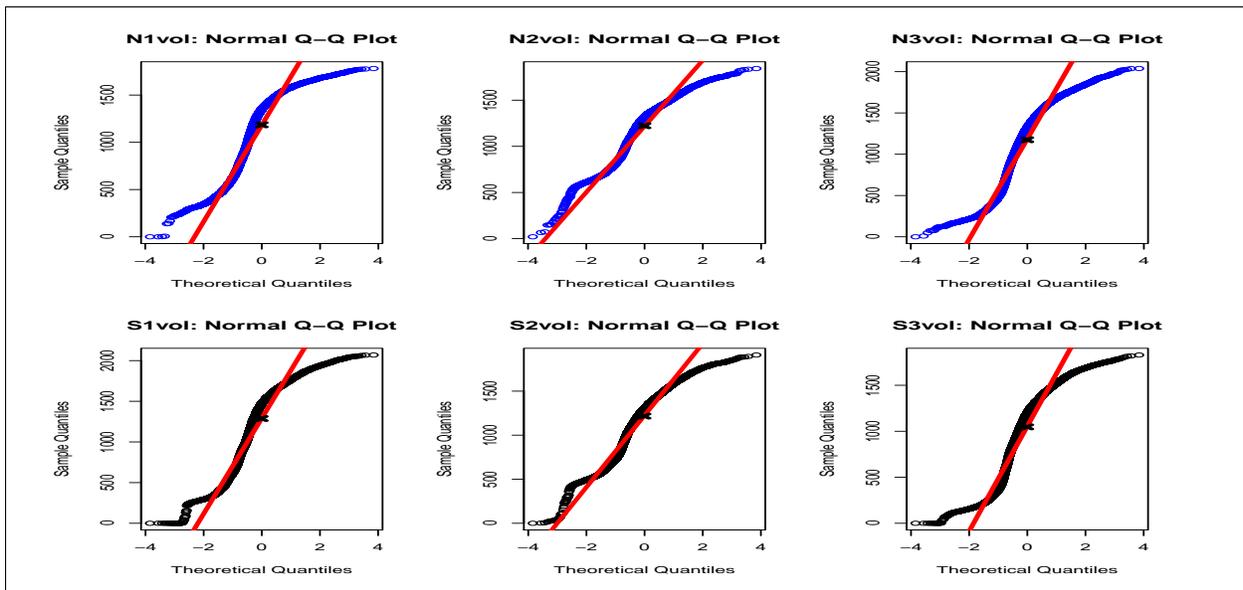
Figure 6: Individual Lane-Volume Normal Q-Q Plots. If traffic volume were normally distributed, the Q-Q plots would follow the straight lines. The "X"'s on the lines mark the volume means. The position of the "X"'s high in the plots indicates skewness.

19

4.2.3 Sample Time Series: One Week Southbound

Summary statistics do not capture patterns revealed in time series. Figure 7 shows three time series for traffic volumes on the three southbound lanes for the week (7 days, 168 hours) beginning midnight on 30 April 2008 and ending midnight 4 May 2008. The volume for Lane 2 is stacked on the volume for Lane 1. Volume for Lane 3 is stacked on the other two so the contour for Lane 3 also represents the total southbound volume per hour.

The most notable revelation of Figure 7 is the marked diurnal variation in traffic volume which is due to heavy traffic during the daytime and light traffic after midnight and before dawn. The "notched" structure within the peaks can be attributed to day-time subpeaks during morning and evening rush hours with still heavy but lessor traffic midday. Also notice that the undulation is not sinusoidal but is asymmetric: The daytime highs are broad (with subpeaks and notches) whereas the pre-dawn valleys are short and "V"-shaped.
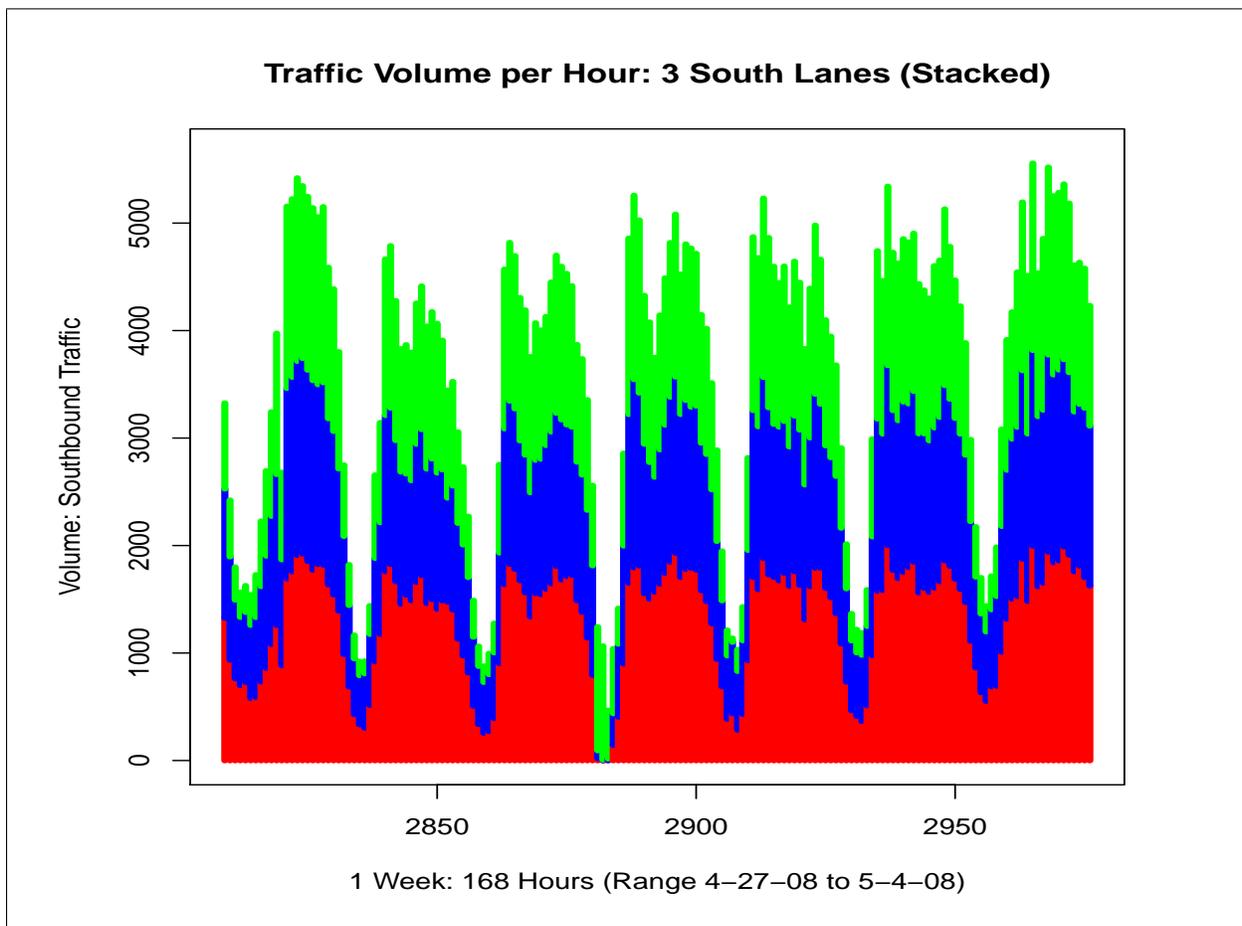
Figure 7: Hourly volume, southbound lanes, week of 4-27-2008 to 5-04-2008. Volume for middle lane stacked onto volume for rightmost lane. volume for leftmost lane further stacked onto that of middle lane.

4.3 Multiple Lane Relationships: Fractional Volumes

So far, the six traffic lanes have been treated as if their volumes were independent. But traffic on one lane is not totally unrelated to traffic on the other two lanes in the same direction. For example, an accident in one lane can affect traffic in the other two same-direction lanes. An example of this is in Figure 7 where, shortly after midnight on April 30, 2008, the depressed volume for several hours in the two rightmost southbound lanes, and the corresponding increased volume on the leftmost lane, suggests that an accident occurred. *There is more discussion of this conjectured accident later.* It is also possible that an accident in one direction can affect traffic in the opposite direction, for example, by inducing rubber-necking. And of course, severe weather can depress traffic across all lanes and directions simultaneously.

One way to assess multiple lane relationships is to look at lane traffic as a percent of total multi-lane volume. Another way is to look at the correlations and scatterplots.

4.3.1 Lane Volume as a Percentage of Multi-Lane Volume

Raw traffic counts do not readily convey relative traffic patterns across lanes. A priori, one might expect each southbound and each northbound lane to carry 1/3rd of the volume in each direction. Table 2 presents basic statistics for each lane's volume as a percent of the total volume in the same direction. The mean and median percent lane volumes are very roughly 1/3rd, but there is a wide range. Volume in both the south and north middle lanes (S2 & N2) never exceeded 2/3rd's of the total direction volume, whereas on at least two (hourly) occasions, the volume on the two other southbound lanes was almost 100%.

Table 2: Lane Volume as Percent of Total Volume in One Direction

	S1	S2	S3	N3	N2	N1
Minimum	0	0	0	0	6.00	0
Mean	35.95	36.10	27.95	31.23	35.84	32.92
Median	36.52	33.52	29.80	33.15	33.52	32.70
Maximum	99.90	65.10	99.44	53.07	61.70	90.30

The wide range in percent lane-volumes is most easily seen in the boxplots in Figure 8. Figure 8 also shows that the interquartile ranges are relatively tight as are the whisker extents. One consequence of the relative tightness is that there are a large number of outliers. The pattern of tight ranges with many outliers in Figure 8 stands in sharp contrast to the pattern of broad ranges and few outliers among the (raw) lane volumes shown in Figure 4.[3]

---

[3]Since the two figures are drawn from the same data, why is the information conveyed so different? The reason is that raw and percent volumes are not mere transforms of each other. Besides the difference in ranges and numbers of outliers between Tables 1 and 2 and Figures 4 and 8, there is a more fundamental difference between volume and percent-volume statistics. Volumes for all lanes can have any values whatsoever (the discussion at the start of this section notwithstanding). However, since percentages for traffic in the same direction must total 100%, there are only two degrees of freedom in determining the percentages for three same-direction lanes. That is, percentages capture a relationship among the lanes and in so doing consume one degree of freedom. This loss of a degree of freedom becomes important in selecting a variance-covariance
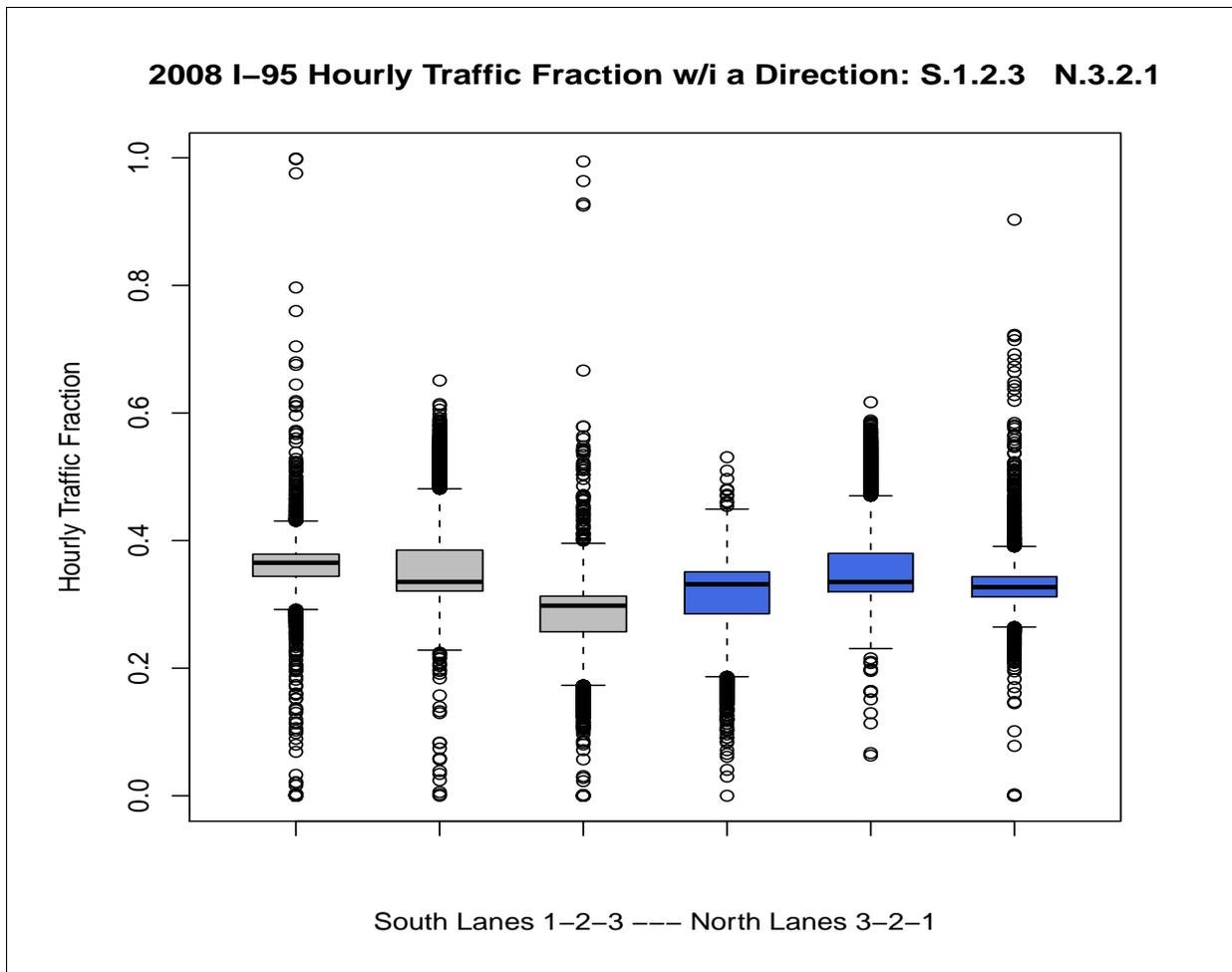
Figure 8: Hourly Traffic: Fractional volume carried by lanes in same direction

4.3.2 Evaluating Normality Assumptions: Fractional Volumes

Since the boxplots in Figures 4 and 8 are so different, and the fact that the same-direction fractional lane-volumes entail one degree of freedom fewer that raw lane-volumes, we would expect that fractional-volume histograms are different from the raw-volume histograms shown in Figure 5. Indeed, the fractional-volume histograms in Figure 9 are very different. The histograms are clearly unimodal in contrast to the bimodal histograms in Figure 5. But the very tall single-mode peaks suggest a leptokurtic distribution different from that associated with a normal distribution. Also, four of the six histograms of Figure 9 show a skewed tail. These two factors suggest checking for non-normality using normal Q-Q plots.

As were the normal Q-Q plots in Figure 6 for lane-volume distributions, the Q-Q plots in Figure 10 are markedly non-linear and highly contorted indicating that the distribution of the fractional lane volumes are clearly not normally distributed. This means that the χ^2 probabilities and ellipses for fractional lane-volumes associated with the (squared) Mahalanobis distance values discussed below are, at best, suggestive.

---

matrix for determining the Mahalanobis distances of the percent volumes.
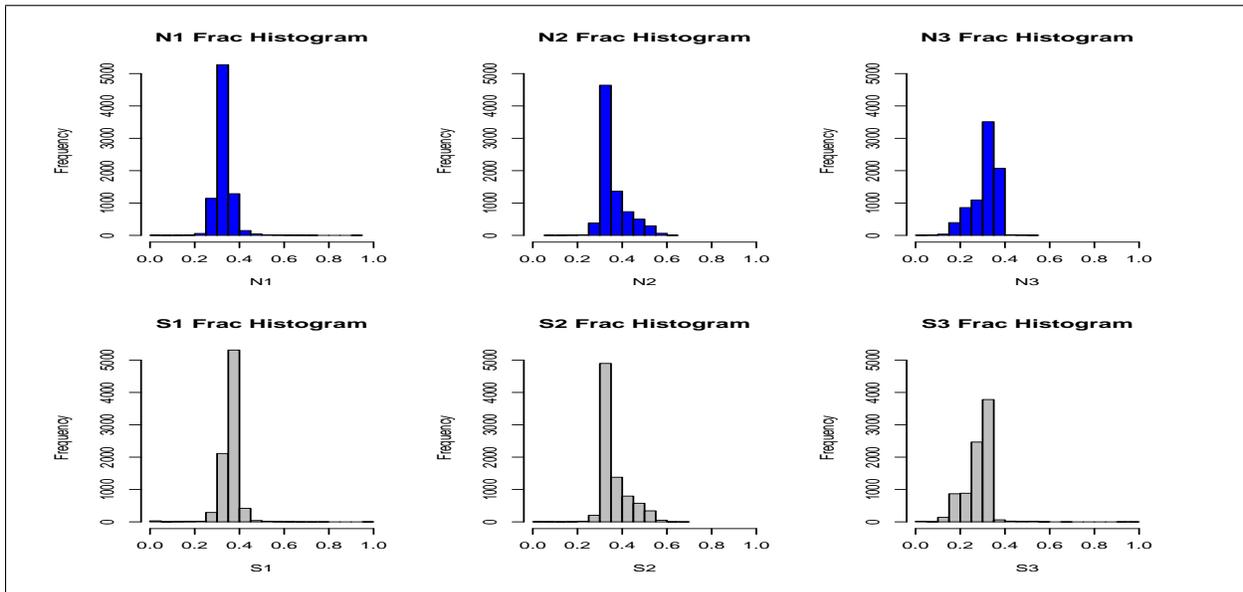
Figure 9: Histograms of fractional volume carried by lanes in same direction
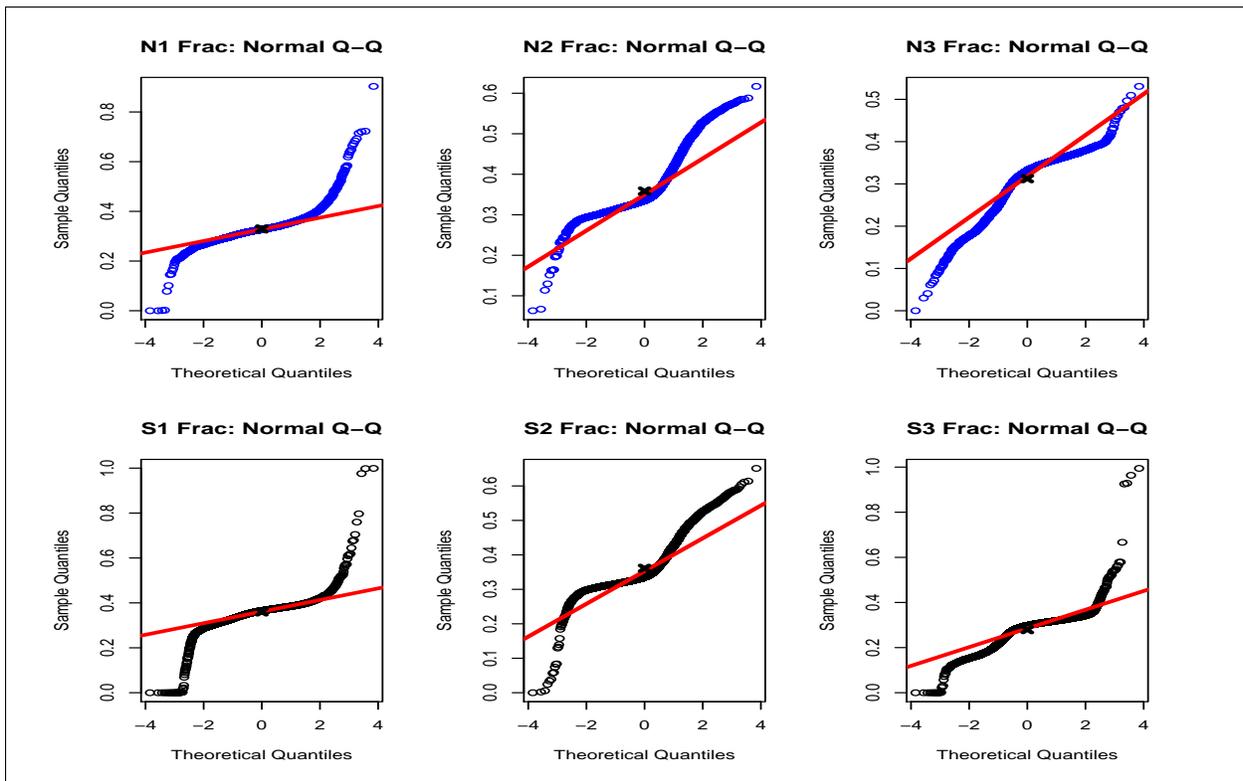
Figure 10: Normal Q-Q plots of fractional volume carried by lanes in same direction. If fractional volumes were normally distributed, the Q-Q plots would follow the straight lines. The "X"'s on the lines mark the fractional-volume means.

4.3.3 Sample Time Series: One Week Southbound Percentages

As an example of the value of looking at percent or fractional volume, Figure 11 shows the hourly fractional volumes for the same one week's traffic shown in Figure 7. The multi-hour anomaly shortly after midnight on April 30, 2008 noted earlier in the discussion of Figure 7 stands out more sharply in Figure 11: For several hours, traffic all but disappears from the two rightmost southbound lanes and the left lane carries almost 100% of the volume. As Figure 7 shows, raw volume was generally habitually low after midnight. By using percentages, the change in pattern stands out.
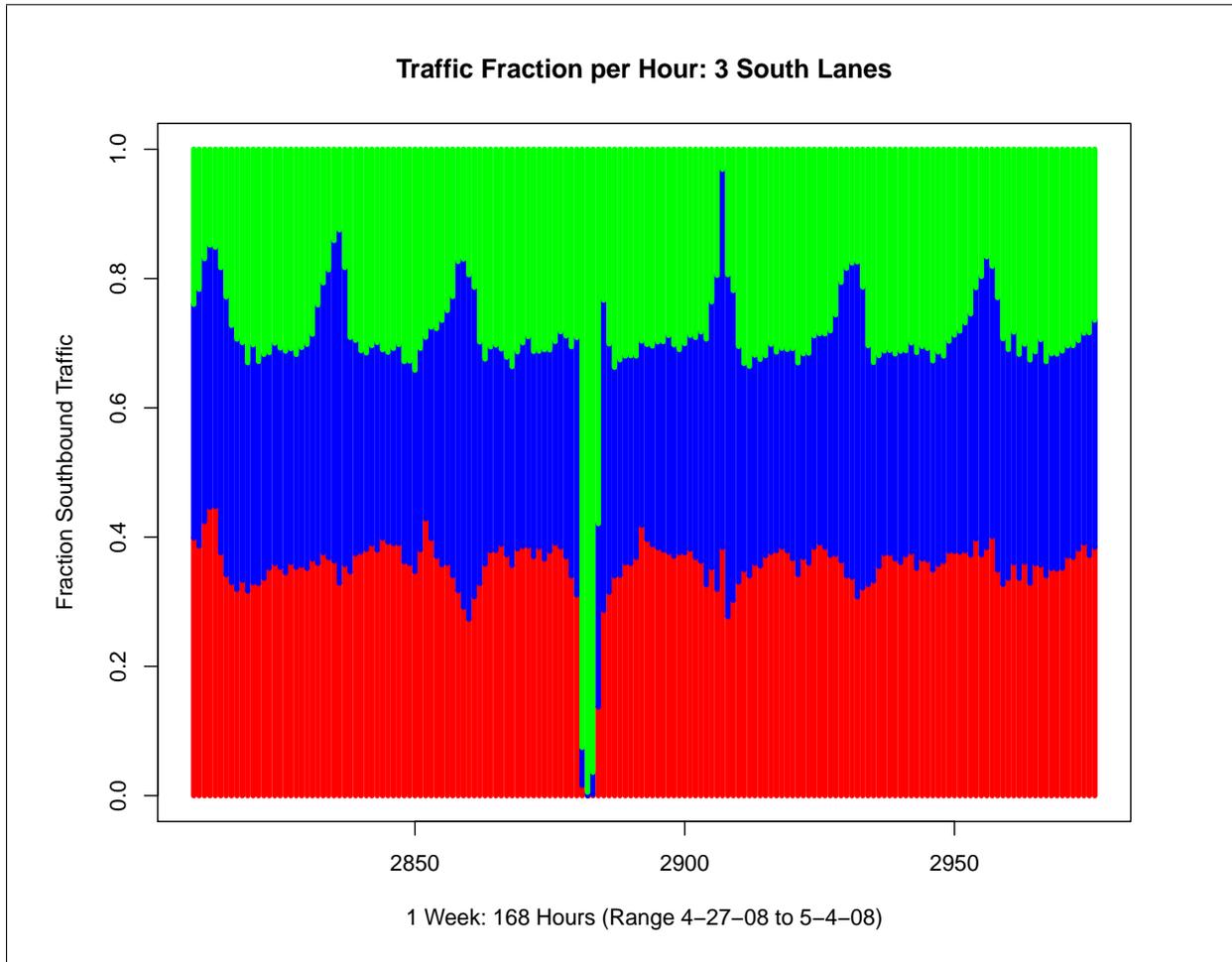
Figure 11: Hourly fractional-volume, southbound lanes, week of 4-27-2008 to 5-04-2008. Fraction for middle lane stacked onto fraction for rightmost lane. Fraction for leftmost lane further stacked onto that of middle lane

One reason for looking at percent-volumes was the expectation that percentages would mitigate, if not eliminate, the diurnal patterns visible in Figure 7: Although rush-hour volume is high and pre-dawn traffic is low, the relative proportion of traffic per lane "should" remain a constant at about 33.3%. Contrary to expectation, Figure 11 shows that the diurnal variation evident in Figure 7 still persists even when the data is plotted as percentages. For raw volume (Figure 7), diurnal variation is easy to explain in terms of high daytime traffic,

especially at rush hours, and fewer drivers out in pre-dawn hours. But, lane variation in percentage traffic is a mystery. Perhaps it is easier to switch lanes late at night while traffic volume is sparse whereas lane-switching is somewhat constrained during bumper-to-bumper rush-hour traffic? If so, this further supports the contention that same-direction lane volumes are not independent.

Figure 12 shows the same data as in Figure 11 but the lane percentages are in their own bands to more easily see the diurnal variation per band.
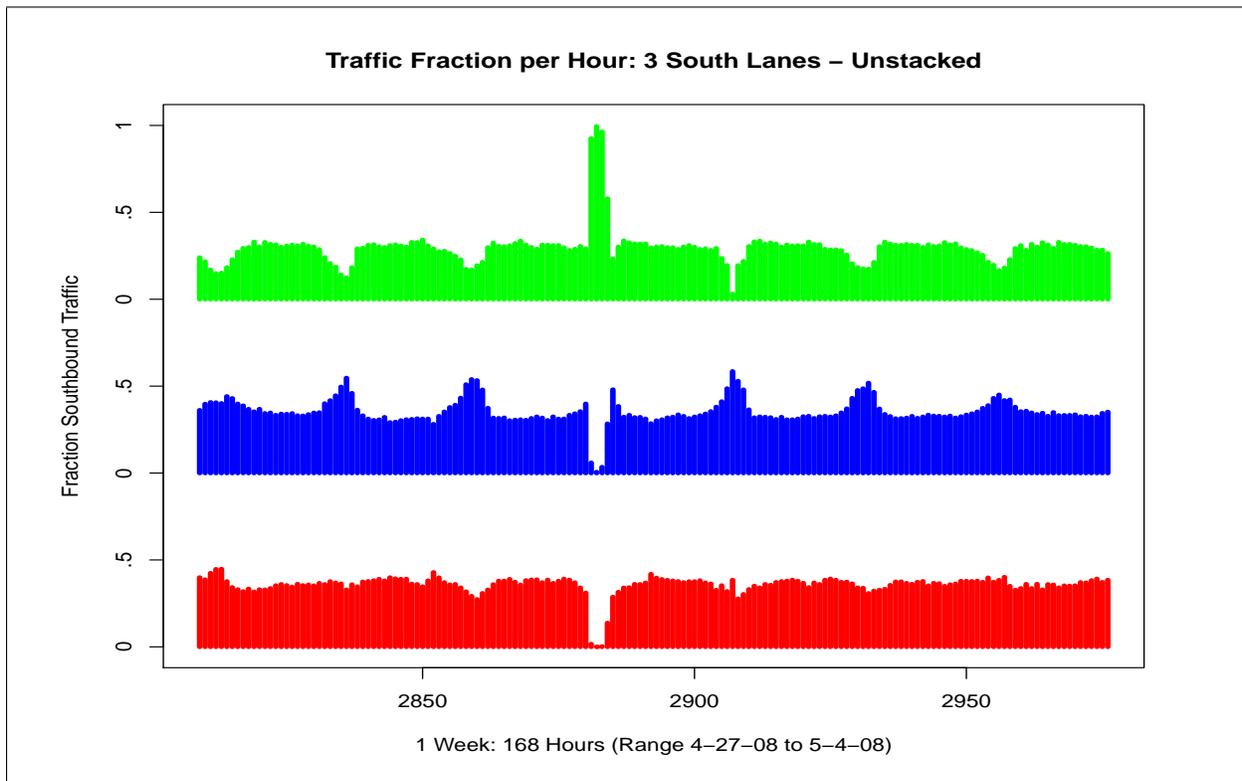
Figure 12: Hourly fractional-volume bands, southbound lanes, week of 4-27-2008 to 5-04-2008.

4.3.4 North & South Percentages: Possible Rubbernecking?

Percentage or fractional volume is also useful for comparing traffic on both directions. On some roads leading into and out of a small city with few alternates routes, morning inbound traffic can exceed morning outbound traffic with a reverse pattern in the evening. However, as the map in Figure 3 shows, I–95 is surrounded by alternate routes and is in a major city, so the northbound/southbound ratio fluctuations are complex as is seen in Figure 13.

Figure 13 spans the same one week shown in Figures 7 and 11 which only showed southbound traffic but which revealed a possible accident on April 30, 2008. What is interesting is that, on a percentage basis, southbound traffic increased relative to northbound traffic during the hours of the (putative) accident. This is shown by the spike in the upper curve of Figure 13. But what is more interesting is the very large spike in the traffic carried by the
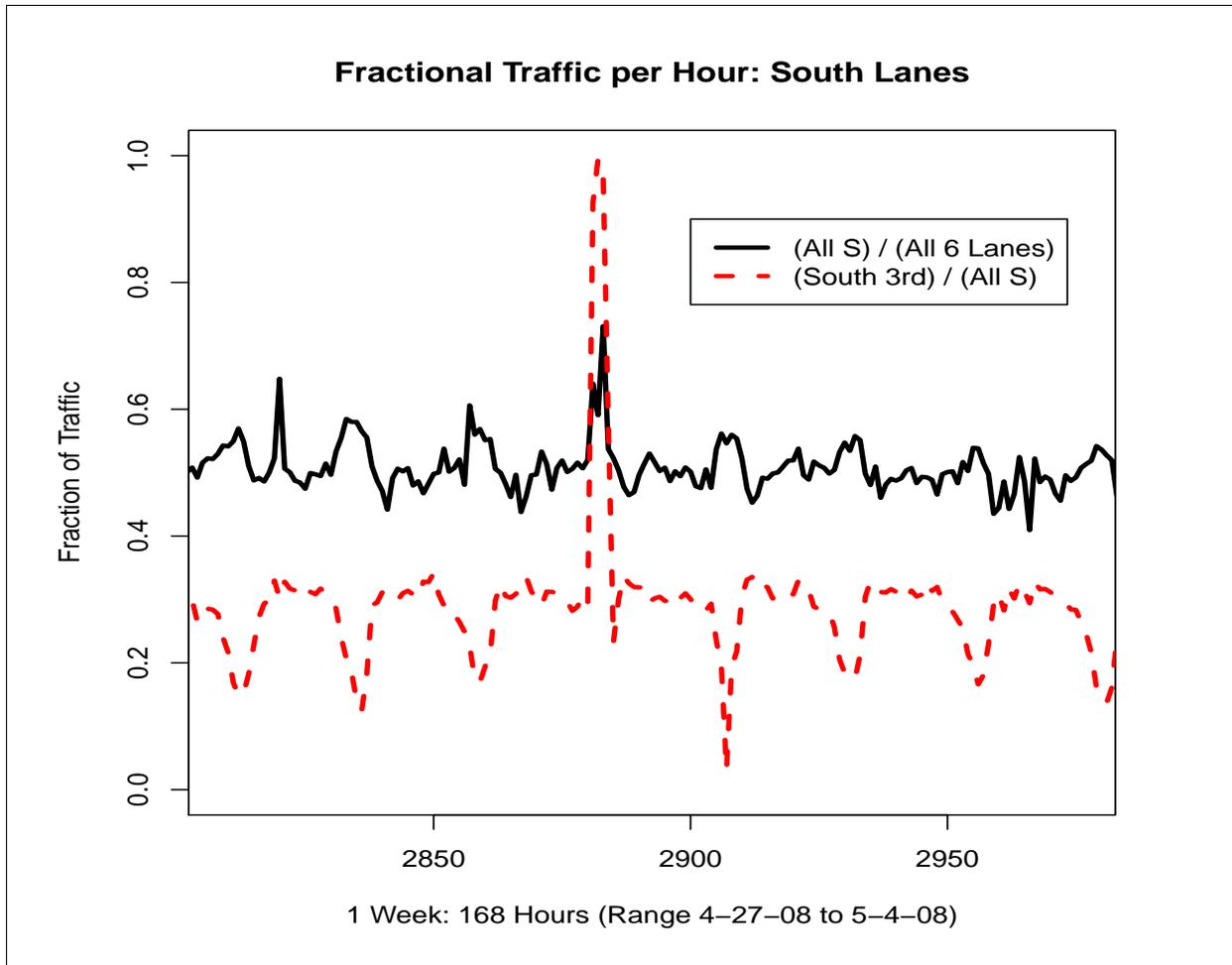
Figure 13: Hourly traffic: Below black line: Fraction carried by all southbound lanes. Above black line: Fraction carried by northbound lanes. Dashed line: Fraction leftmost southbound lane carried of all southbound traffic.

leftmost southbound lane as a fraction of the total southbound volume. We have seen this spike before in Figures 7 and 11, but now it can be seen to be coincident with the north-south spike of April 30.

One possible explanation for the coincidence of the two spikes is that northbound absolute and relative volume decreased as a consequence of rubbernecking: Northbound drivers slowed to observe the accident scene. The putative accident of April 30 can be verified by accident records. The rubber-necking conjecture cannot be verified with this dataset, but the type of spike pattern observed here could be checked using datasets that include traffic speeds as well as volume.

If other data support the rubbernecking conjecture, then we would have a method for indirectly detecting correlated activity. "Correlated" in the sense that there is a direct contemporaneous linkage between two events.

Direct contemporaneous linkages, in general, are easily viewed in scatter plots.

4.4 Multiple Lane Relationships: Scatterplots

In preparation for the Mahalanobis distance analysis, and since the individual lane-volumes are skewed and exhibit a bimodal distribution with a minor mode about 1/4 to 1/3rd the height of the major mode, it is useful to look at the underlying bi-variate distributions as scatterplots.

4.4.1 Multi-Lane Volume Scatterplots

Figure 14 pairs the total northbound hourly volumes with the corresponding total southbound hourly volumes. Paired north and south data is missing for 1,185 of the 8,784 hours in 2008 leaving $n = 7,599$ hours of complete data in both directions. (Southbound data is missing for 502 hours, northbound for 773 hours, and both southbound and northbound simultaneously missing such that fewer than $502 + 773$ hours are missing.) The "top heavy" skew of daytime volume is evidenced by the large dense sub-region of Figure 14. The existence of a minor mode of low-volume nighttime traffic is a also apparent by a slight thinning between the two ends of the deformed ellipse or between the large head and small tail of the pattern.
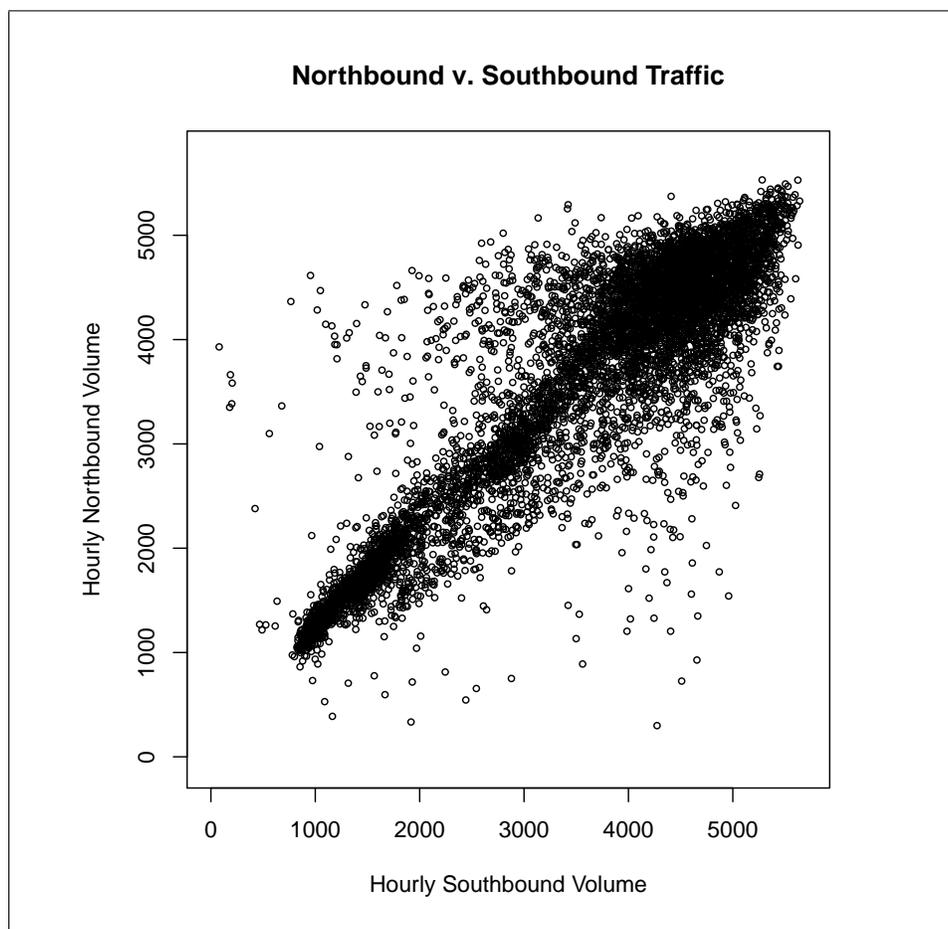
Figure 14: Scattterplot of total northbound vs. total southbound volume. Full year 2008. $n = 7,599$ data points.

Figure 15 shows the volumes for the paired-lane components whose totals appear in Figure 14. The overall lane-to-lane volume correlations and non-unimodal distributions that are evident in Figure 14 are also evident in the components of Figure 15 but not as prominently and with their own individual peculiarities. Note that the ranges are not always equal among the plots.
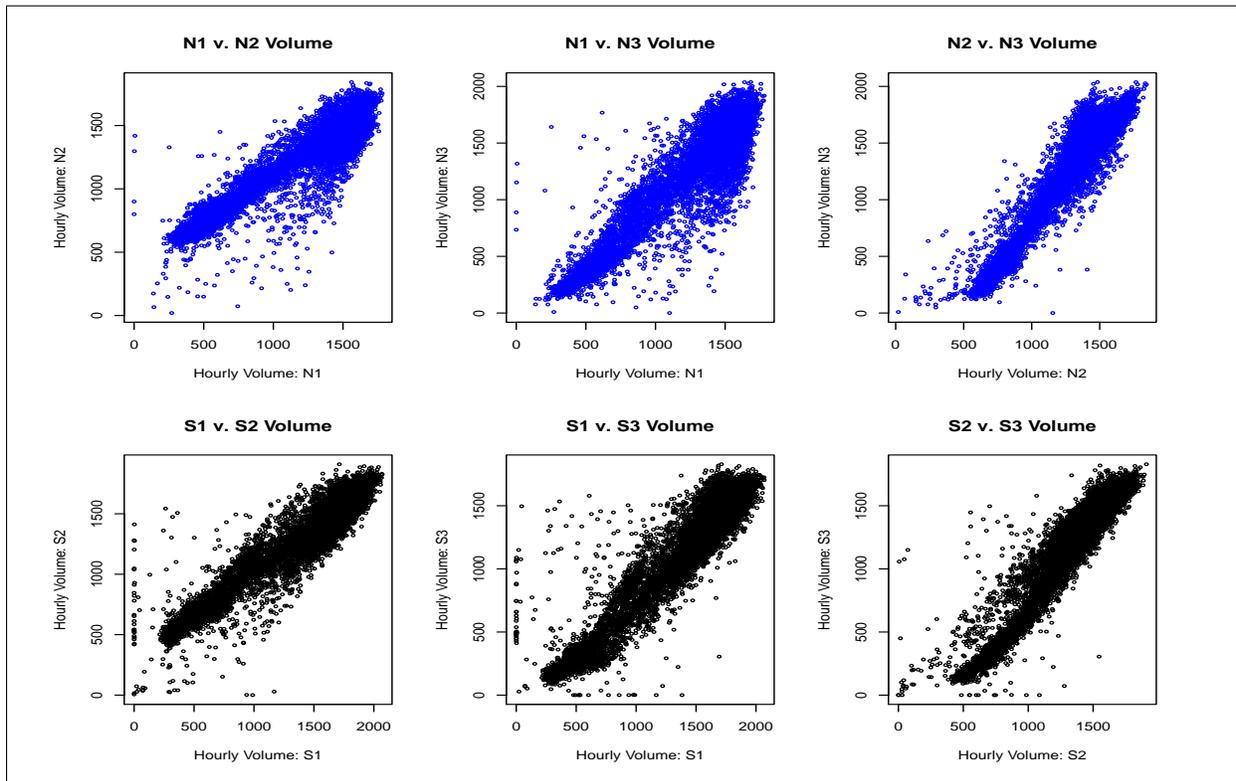
Figure 15: Scatterplots of traffic volume in same-direction lane-pairs. Full year 2008.

4.4.2 Multi-Lane Fractional-Volume Scatterplots

Fractional-volume multi-lane patterns are different from the analogous raw-volume patterns due to the loss of a degree-of-freedom forced by the constraint that the relevant total fractional volume must be unity. Most notably, since raw volume is (usually) unconstrained, high volume in one lane is normally accompanied by high volume in same-direction lanes. However, since percentages are linked, a percentage increase in one area is generally accompanied by a percentage decrease in another area.

The inverse relationship is 100% when comparing traffic in one direction as a percent or fraction of the total volume in both directions. Since fractional northbound volume must be 1 minus the fractional southbound volume, there is only one degree of freedom and the correlation between the the fractional volumes is exactly $r = -1$. A scatterplot analogous to Figure 14 (fractional southbound volume versus fractional northbound volume) would show all points along a straight line with negative slope.

Since there are two degrees of freedom for the fractional volumes among each set of three same-direction lanes, the scatterplots in Figure 16 are more interesting. All clouds show

a clear negative slope reflecting the usual inverse relationship that an increase in fractional volume in one lane is accompanied by a decrease in the other two lanes. A priori, all fractional volumes would be about 1/3, so all centroids would have coordinates (.33, .33). Indeed, the clouds are generally located toward the lower left quadrants of the plots since all axes run from zero to 1 for ease of comparison.

What is most striking is the very poor elliptical structure of the clouds. The less the cloud structure is described by confocal elliptical shells, the less accurately the (squared)-Mahalanobis distance of the points follows a χ^2 distribution. We turn now to the Mahalanobis analysis.
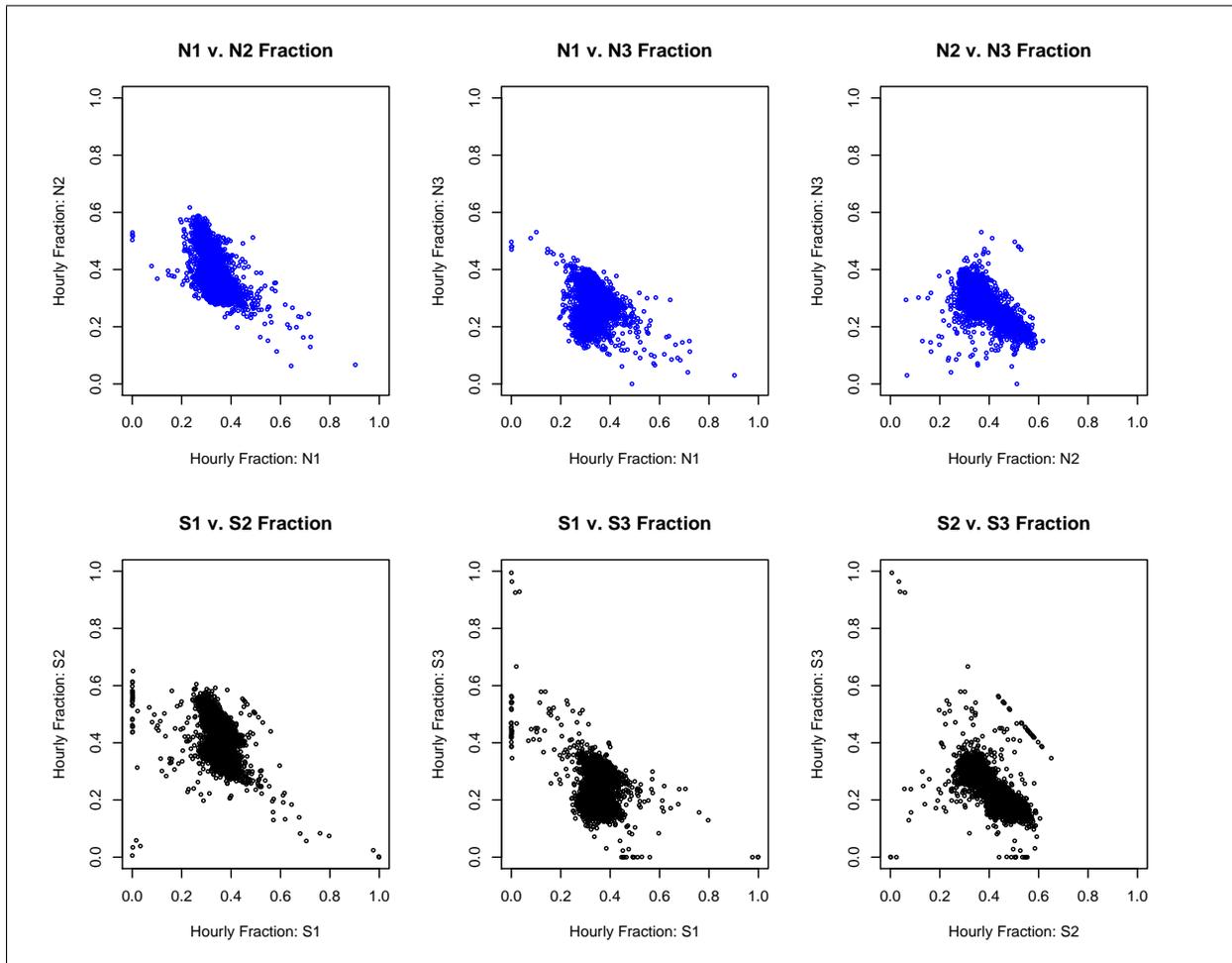
Figure 16: Lane-lane fractional same-direction volume in same-direction lane-pairs. Full year 2008. All ranges are 0 to 1.

5 MAHALANOBIS DISTANCE (d_m) RESULTS

Mahalanobis distances are a nice way to reduce the dimensionality of multivariate data sets. Since a multivariate data set can be conceived as points in a multivariate space or scatterplot, the several elements of one multivariate data point can be associated with the data-point's (Mahalanobis) distance from the centroid of the scatterplot.

We report here the Mahalanobis distance for each triplet of the per-lane hourly-volumes in each direction. Since the raw volumes are unconstrained, there are three degrees of freedom with each volume triplet, and three degrees of freedom for comparing the (squared)-Mahalanobis distances with the χ^2 distribution.

We also report the Mahalanobis distance for the same data transformed into fractional traffic per lane per direction since the previous section showed that the raw volumes and fractional volumes are not perfectly correlated and hence convey different information. However, since each fractional-volume triplet must sum to one, there are only two degrees of freedom for each fractional-volume triplet, and thus only two degrees of freedom for comparing these (squared)-Mahalanobis distances with the χ^2 distribution.

Dimensionality reduction is not enough: We are searching for unusual hourly patterns and that can be done in several ways besides using χ^2.

5.1 Detecting Unusual Patterns Using d_m

The mean percent southbound volume carried by each of the three southbound lanes was 36%, 36%, and 28% respectively (See Table 2). When actual lane-volumes per hour are plotted with lines between adjacent lane-volumes, the resulting patterns are most often similar to the sample patterns shown in Figure 17. The several lines are progressively three hours apart on January 1, 2008. They are intended to convey, in a static figure, that hourly lane-volumes vary systematically throughout the day.

When the linked lines are shown as a movie with one hour's three volumes and linked lines per frame, the resulting impression is that of a bird in flight rhythmically ascending and descending with wings flapping as in Figure 17 most of the time. The rhythmically ascending and descending of the linked lines corresponds to the normal diurnal traffic ebbs and flows depicted in Figure 7.

However, every now and then, the wings assume an unusual attitude defined by a rare combinations of volumes, say, 1000, 1500, and 5 vehicles per hour or a rare percentage pattern, say, 90%, 9%, and 1%. Both of these patterns would suggest an accident in the third lane.

These examples show how the multivariate data points can be conceived of as geometric figures or patterns. Mahalanobis distance can also serve as a means for determining typicality and unusualness among sets of geometric figures.
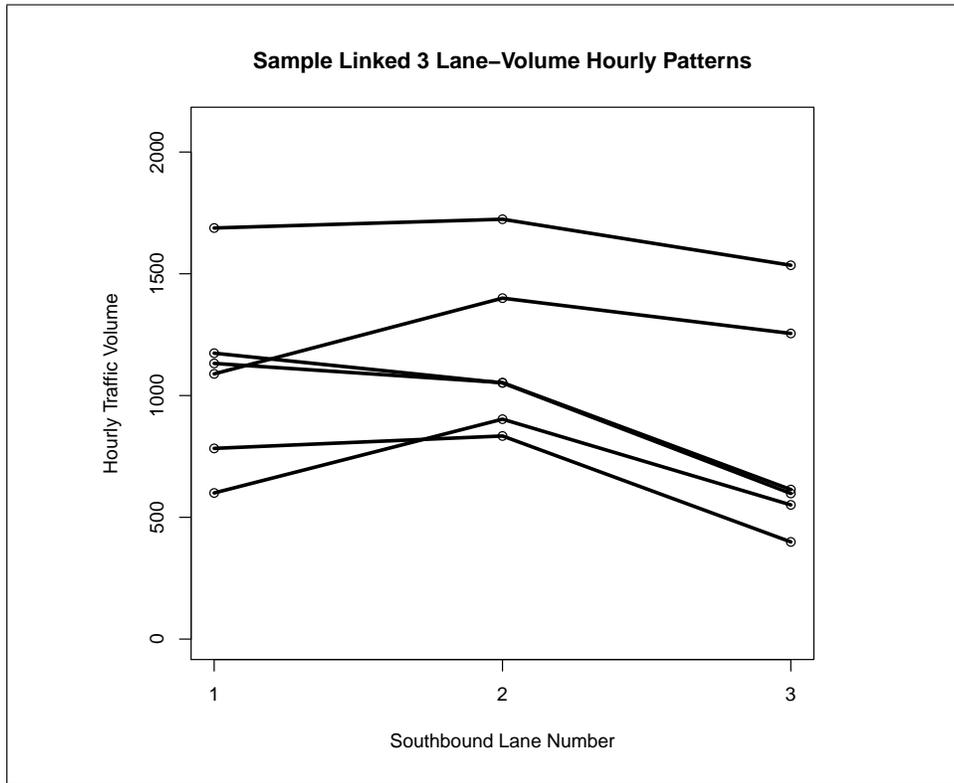
Figure 17: Sample linked 3-lane hourly volumes. Linked hourly samples are progressively three hours apart on January 1, 2008. The varying heights reflects the change in traffic throughout the night. If shown as a movie, the hourly changes resemble a bird in flight.

5.2 Full Year 2008 Fractional & Raw Volume d_m

Typical and unusual Mahalanobis distances for hourly volumes and fractional-volumes per traffic direction are captured in Table 3 and Figures 18 and 19.

Table 3 lists basic Mahalanobis distance statistics: Minimums, means, medians, and maximums. No minimum d_m is exactly zero since such no multivariate data point was exactly equal to its respective centroid. That the medians are smaller than the means suggests a high-end skew. For both directions, the maximum d_m for fractional volume exceeds that for raw volume. This is peculiar and suggest looking at the histograms—which we will do after looking at the entire year's d_m profiles.

Table 3: Mahalanobis-Distance Volume and Fractional-Volume Statistics per Traffic Direction

	S123.Vol	S12.Frac	N123.Vol	N12.Frac
Minimum	0.08	0.02	0.07	0.01
Mean	1.54	1.06	1.56	1.13
Median	1.45	0.79	1.46	0.90
Maximum	12.14	14.80	8.80	15.36

One speculation for the difference in the ranges of the Mahalanobis distances when using volume versus fractional-volume data is that fractional-volume filters out, in some sense, variation due merely to diurnal variation. For example, consider three 3-lane raw volumes of [100, 100, 100], [500, 500, 500], and [1000, 1000, 1000] vehicles as might possibly occur during the transition from pre-dawn to post-dawn traffic. The volume-based d_m values for these three cases would all be different, whereas the fractional-volume-based d_m values would all be the same since in all three cases the fractional volume pattern is [.33, .33, .33]. On a relative basis, this suggests that there is less variation in the fractional-volume patterns than in the raw-volume patterns. Since the procedure for Mahalanobis distance, in essence, weighs scores inversely by their standard deviations, the *relatively* smaller variation in fractional-volume patterns might produce larger Mahalanobis distances.

Figures 18 and 19 show Mahalanobis distances for volume and fractional volume on southbound lanes and northbound lanes, respectively. The gaps in both Figures are due to missing data. The dark bands suggest that the vast majority of Mahalanobis distances are less than four. Outliers several times the mean and median values are rare and seem to be somewhat correlated within a direction. That is, an outlier signaled by a volume data d_m often is also usually signaled by a fractional-volume d_m. Note that the d_m ranges are different within and between traffic directions.
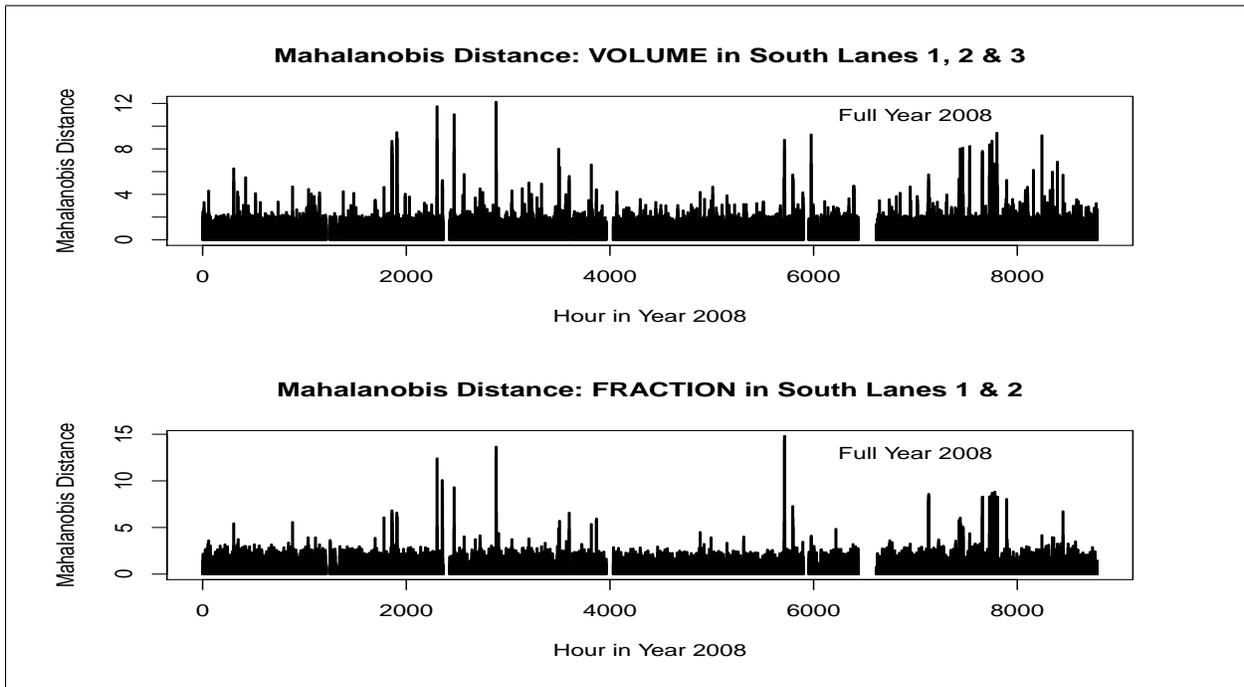
Figure 18: Mahalanobis Distances for Volume and Fractional Volume on Southbound Lanes
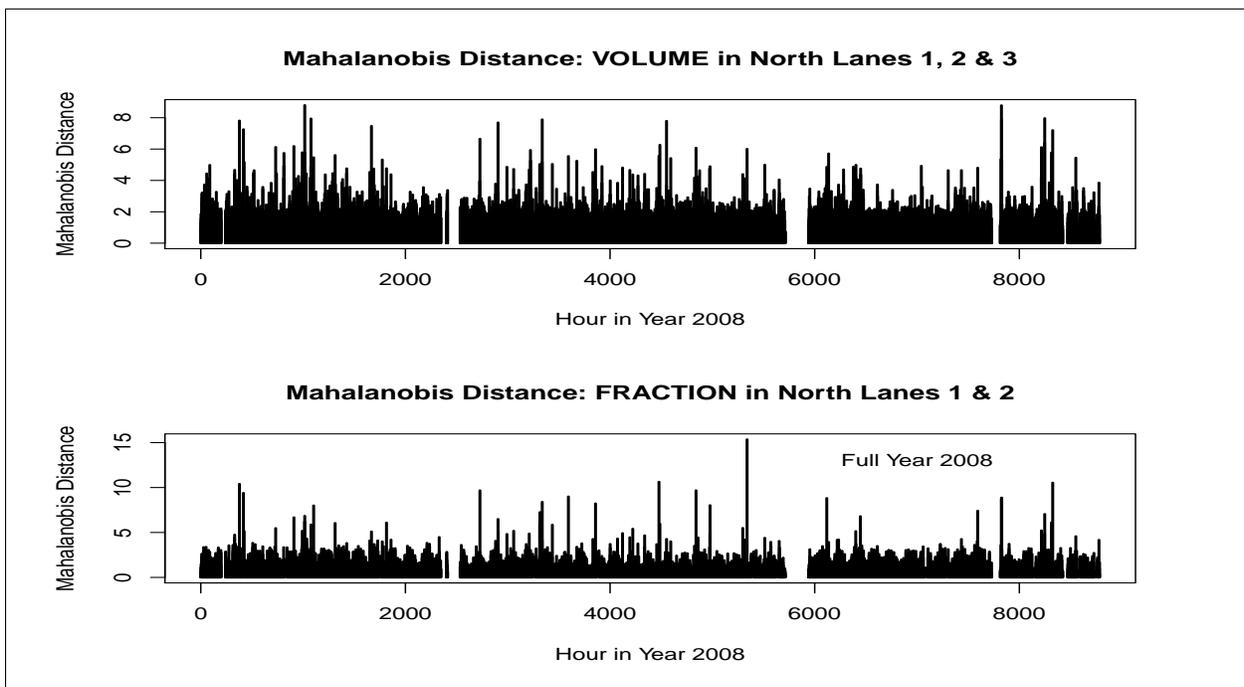
Figure 19: Mahalanobis Distances for Volume and Fractional Volume on Northbound Lanes

5.3   One Week's Fractional & Raw Volume d_m

Figures 18 and 19 each contain 1 year's or 8,784 hour's results in each sub-plot (where results can include a gap for missing data). Graphically, this forces a loss of detail due to horizontal compression: Valleys or troughs are squeezed into oblivion. To enable valleys to be seen, Figures 20 and 21 limit themselves to the 168 hours of the week of April 27 to May 5, 2008. Previous figures (Figures 11, 12, & 13) have suggested a possible accident affecting southbound traffic for several hours shortly after midnight on April 30, and now Figure 21 clearly shows a cluster of very high d_m values at those times. The d_m spike cluster is virtually shouting that something very anomalous was happening during those hours. A large spike is also present that same week in the northbound lanes. Overall, the vertical bars (d_m values) look generally higher on the northbound plots, but that is an artifact due to the difference in the vertical scales.
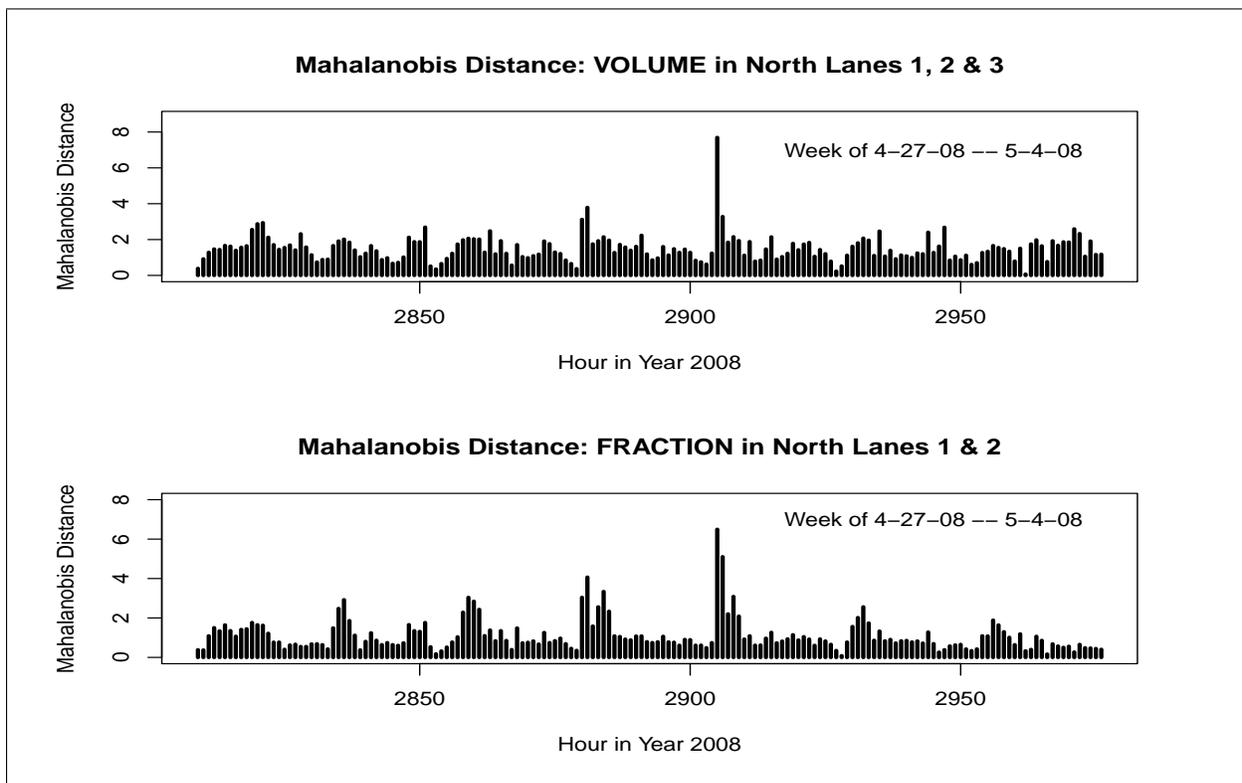
Figure 20: Mahalanobis Distances for Volume and Fractional Volume on Northbound Lanes. Week of 4-27-08: 168 Hours
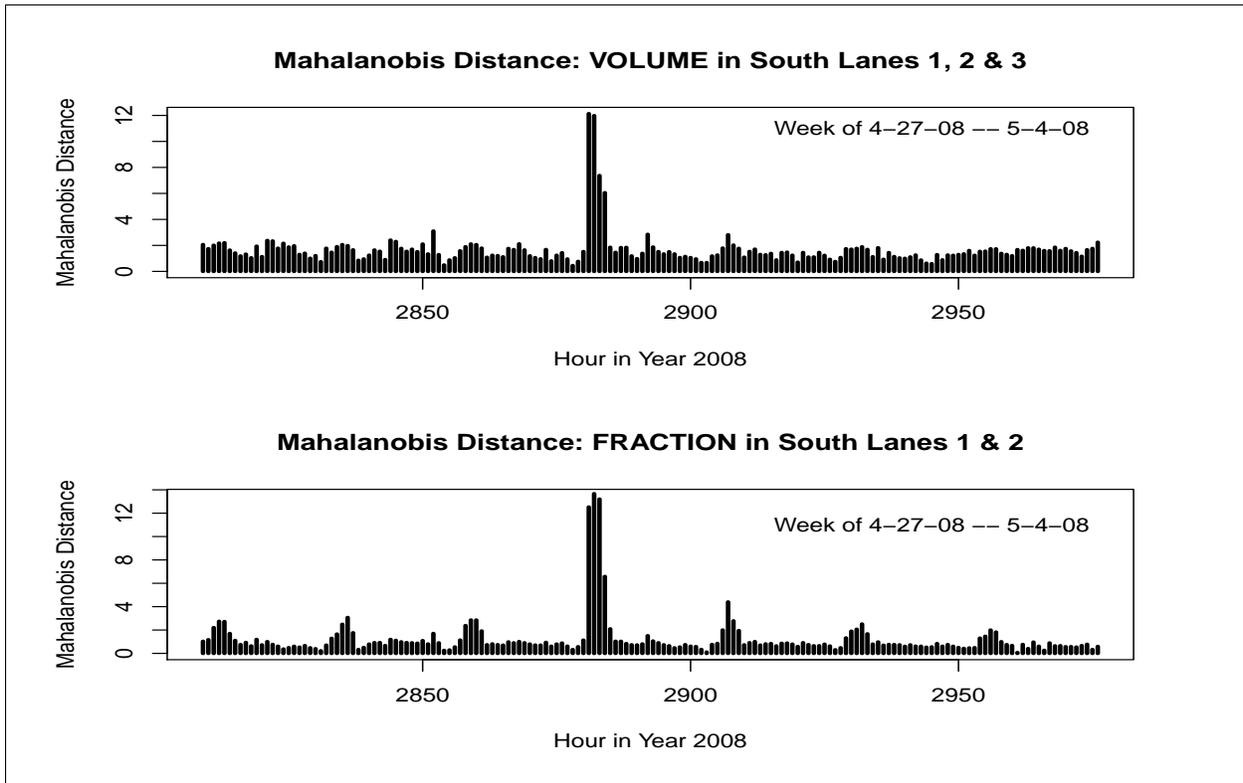
Figure 21: Mahalanobis Distances for Volume and Fractional Volume on Southbound Lanes. Week of 4-27-08: 168 Hours

5.4 Fractional-Volume d_m: Correlation With Volume d_m

In the discussion of Figures 18 and 19, we said that outliers several times the mean and median values are rare and seem to be somewhat correlated within a direction. That is, an outlier signaled by a volume data d_m often is also signaled by a fractional-volume d_m although the d_m ranges are different within and between traffic directions. This conclusion is amply supported in Figure 22: if an hour is an outlier based on the pattern of its three volume components (in one direction), it is also an outlier based on the pattern in *two* of its three fractional-volume components (in the same direction).

Each scatterplot in Figure 22 contains over 8,000 points. Hence, the lower-left corners contain an overwhelming number of points and the relatively few points outside of that corner are indeed outliers on a percentage basis. (1% of 8,000 is 80 – and there are fewer than 80 points discernible outside of the dense cloud in each sub-figure.)
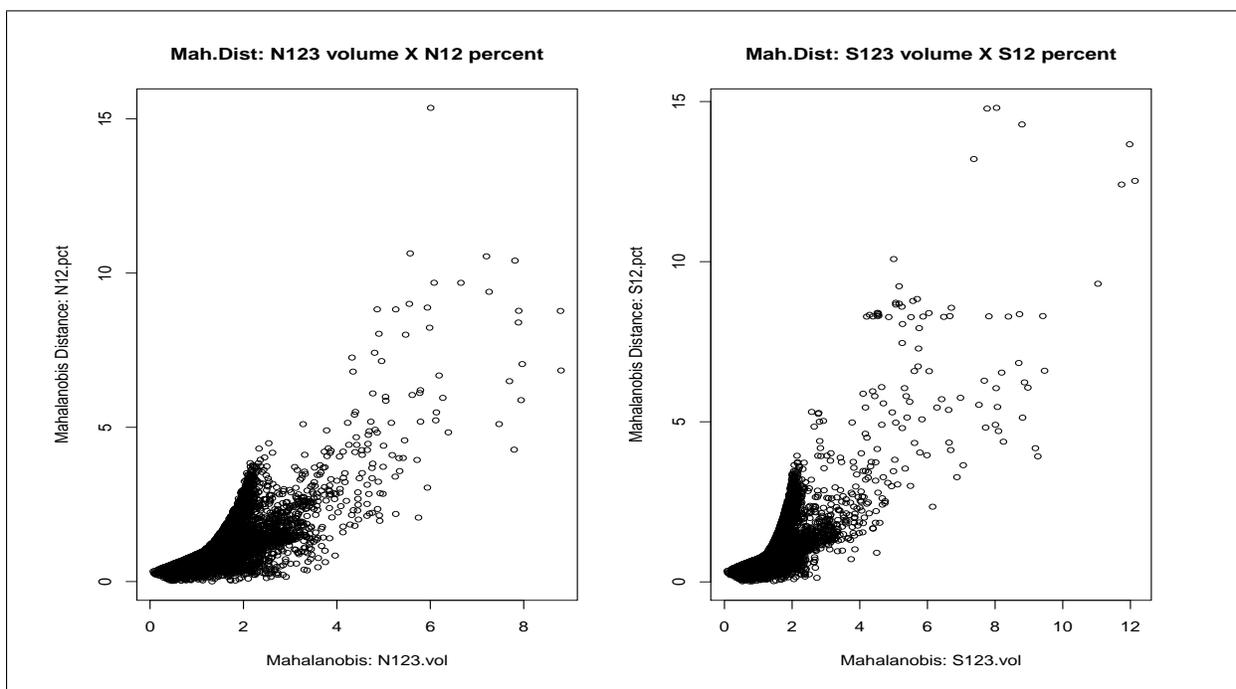
Figure 22: Mahalanobis Distances for Volume vs. Fractional Volume. Left Panel: Northbound Traffic. Right Panel: Southbound Traffic. Northbound and Southbound Volume Ranges are Different

5.5 Fractional-Volume d_m: Correlation Within Directions

We continue the anomaly analysis now using only fractional-volume results within a traffic direction.

This is justified by three facts: (1) Figure 22 showed that an outlier flagged by volume data within a direction is also flagged by fractional-volume data. (2) Mahalanobis distance for hourly volumes within a direction entails three degrees of freedom and requires a three-dimensional ellipsoid to portray all the volume relationships. And (3), Mahalanobis distance

for fractional (or percent) hourly volumes within a direction only entails two degrees of freedom and a two-dimensional scatterplot to portray all the percent-volume relationships.

In particular, we may use any top-row figure in Figure 16 (and its two sets of per-lane fractional-volumes) as the basis for a full analysis of the three northbound lanes, and likewise for southbound traffic.

Figure 23 is a scatterplot of Southbound Lane 1 versus Southbound Lane 2 hourly fractional-volume. These two values contain all the fractional-volume information since the fractional-volume for Lane 3 must be equal to one minus the sum of the two shown lanes fractional-volumes. The unit Mahalanobis distance ellipse as well as the ellipses enveloping 95% and 99% of all points (assuming the data is multivariate normal) are superposed on the scatterplot. The ellipses do not appear to follow the negative diagonal followed by the broad sweep of the points when the far corners are included. But this is an artifact: The ellipses do align with the more vertical-leaning dense cloud of containing most of the over 8,000 points in the plot. It is immediately apparent that any hours containing a 0% or near–0% fractional-volume are outliers.
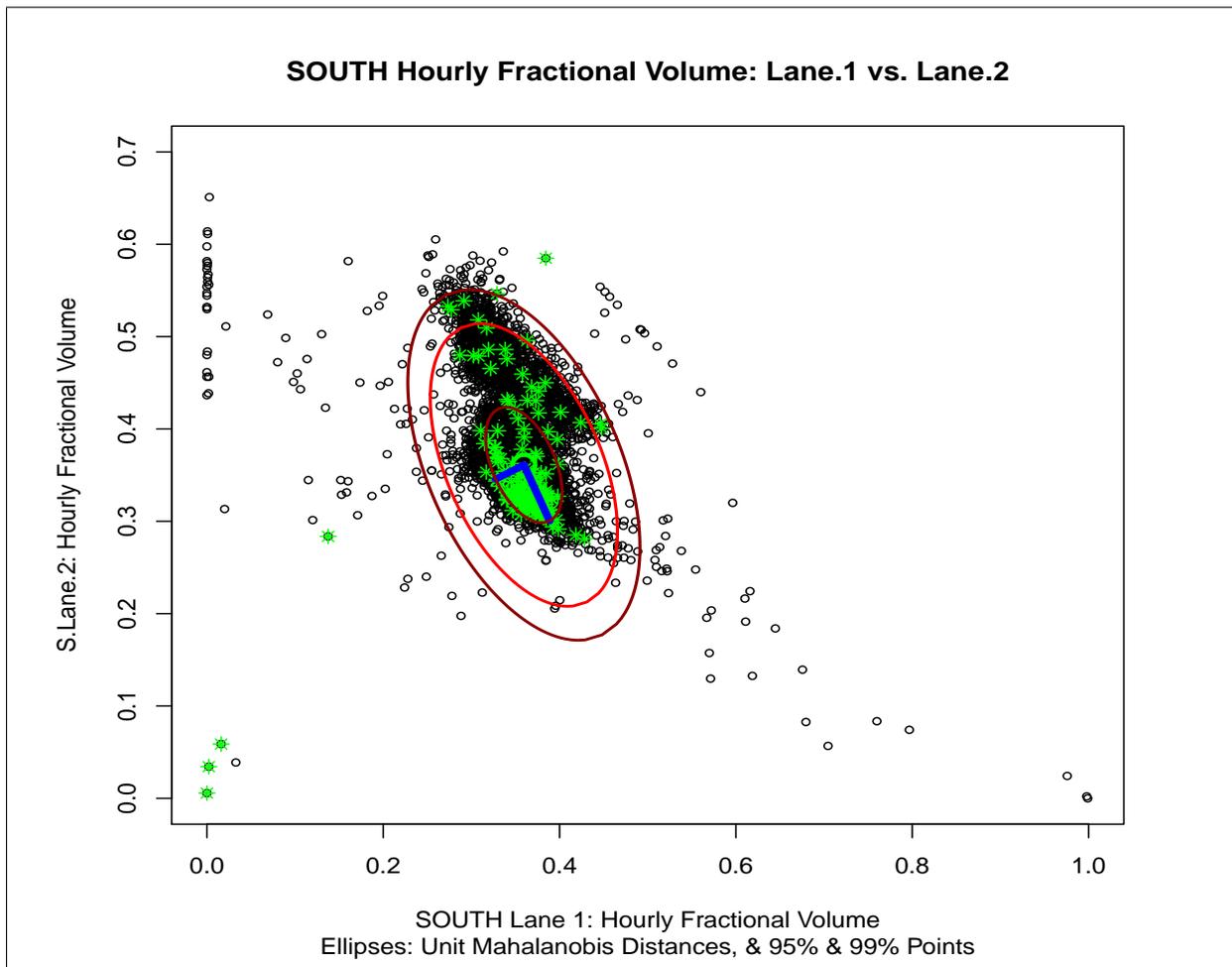
Figure 23: Mahalanobis Distances for South Lane 1 and Lane 2 Hourly Fractional-Volume. Ellipses at $d_m = 1, 2.45, \& 3.03$. The latter two encompass 95% and 99% of all points if data is multivariate normal. The asterisks are the d_m from the week of 27 April, 2008.

Also superposed on the scatterplot are the 168 points from the week of April 27, 2008. These are shown in a lighter color and with asterisks to help them stand out. The vast bulk of that week's fractional-volumes are within the 95% ellipse. What is most interesting are the three points in the lower-left corner which indicate zero or near-zero fractional volumes. Also interesting is the asterisk at roughly (.45,.59) since this point is a middle-point for Lane 1 and is an outlier only due to its unusual pattern when the other lanes are considered. A check with Figure 21 shows that these four points, which do not form a tight cluster in Figure 23, do correspond to the spike-cluster of d_m in the wee hours of April 30. That is, all four points belong to the same hypothesized multi-hour accident.

Figure 24 is similar to Figure 23 but using the fractional-volume data from the north-bound three lanes. The asterisks again correspond to the week of 27 April. Outliers for that week can be compared with the d_m spikes in Figure 20.
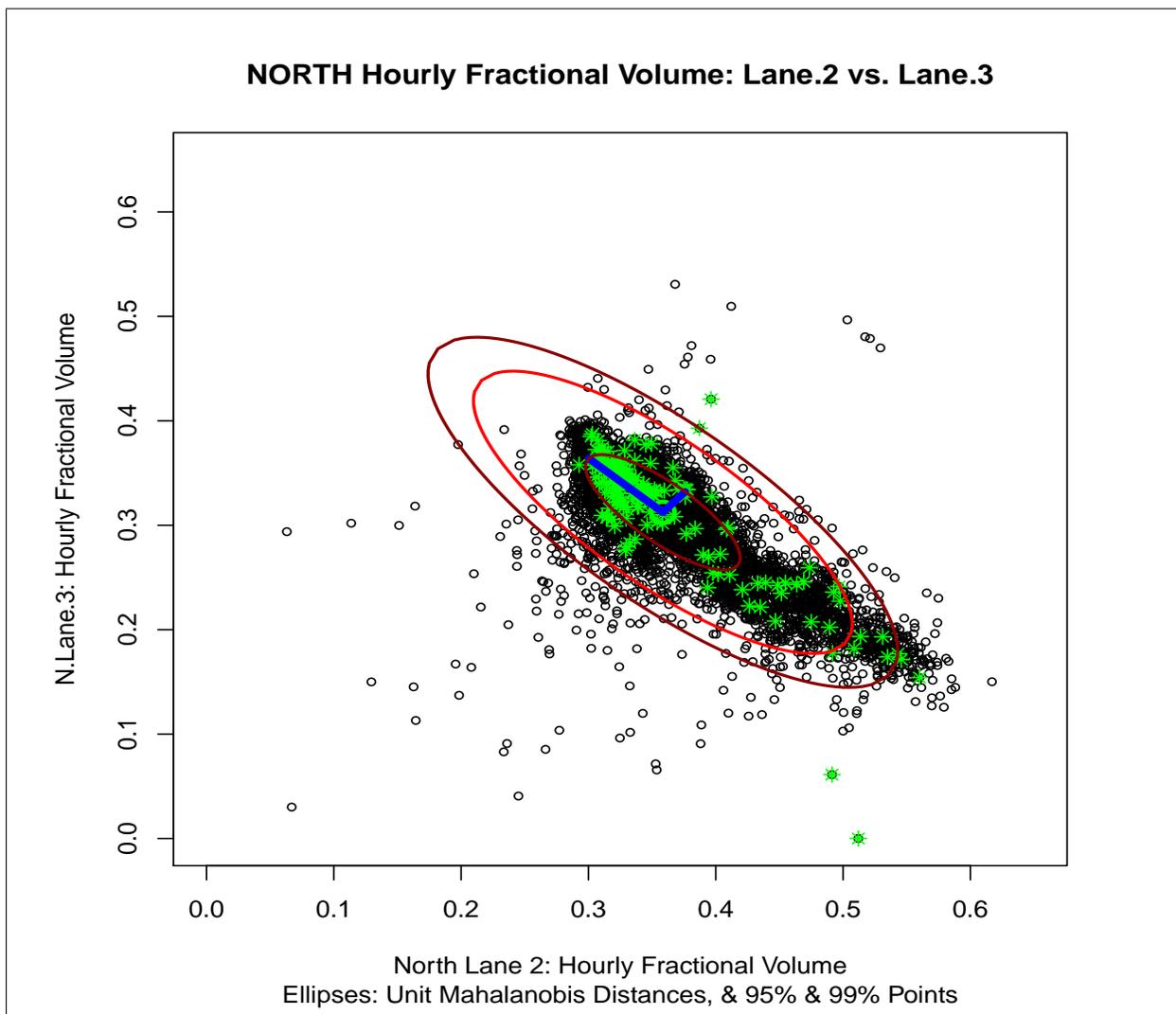

Figure 24: Mahalanobis Distances for North Lane 2 and Lane 3 Hourly Fractional-Volume. Ellipses at $d_m = 1, 2.45, \& 3.03$. The latter two encompass 95% and 99% of all points if data is multivariate normal. The asterisks are the d_m from the week of 27 April, 2008.

5.6 Fractional-Volume d_m^2 and χ^2 Distributions

The size of the ellipses in Figures 23 and 24 are predicated on the assumption that the underlying distributions of squared-Mahalanobis distances are multivariate normal. If that is correct, the d_m^2 should follow a χ^2 distribution as discussed earlier. So, rather than test the empirical and theoretical distributions with quantile-quantile plots, we plotted the χ^2 distribution on top of the empirical d_m^2 histogram. By ensuring that both distributions have unit area, goodness-of-fit can be judged visually. Figure 25 shows the superposed distributions for the d_m^2 of the fractional-volumes on southbound Lanes 1 and 2 and a χ^2 distribution with two degrees of freedom in the left panel and the counterpart using the raw volumes on the three southbound lanes and a χ^2 distribution with three degrees of freedom in the right panel. To aid visualization, the x-axis is clipped at 16 in both panels since the few cases of a d_m^2 greater than 16 do not have a density-height noticeable beyond a slight thickening of the axis. Although there are some overshoots and undershoots between the curves, they appear to cancel each other such that the two curves run fairly on top of each other for higher values of d_m^2 and χ^2. Two vertical bars mark the critical outer .05 and .01 values of χ^2.
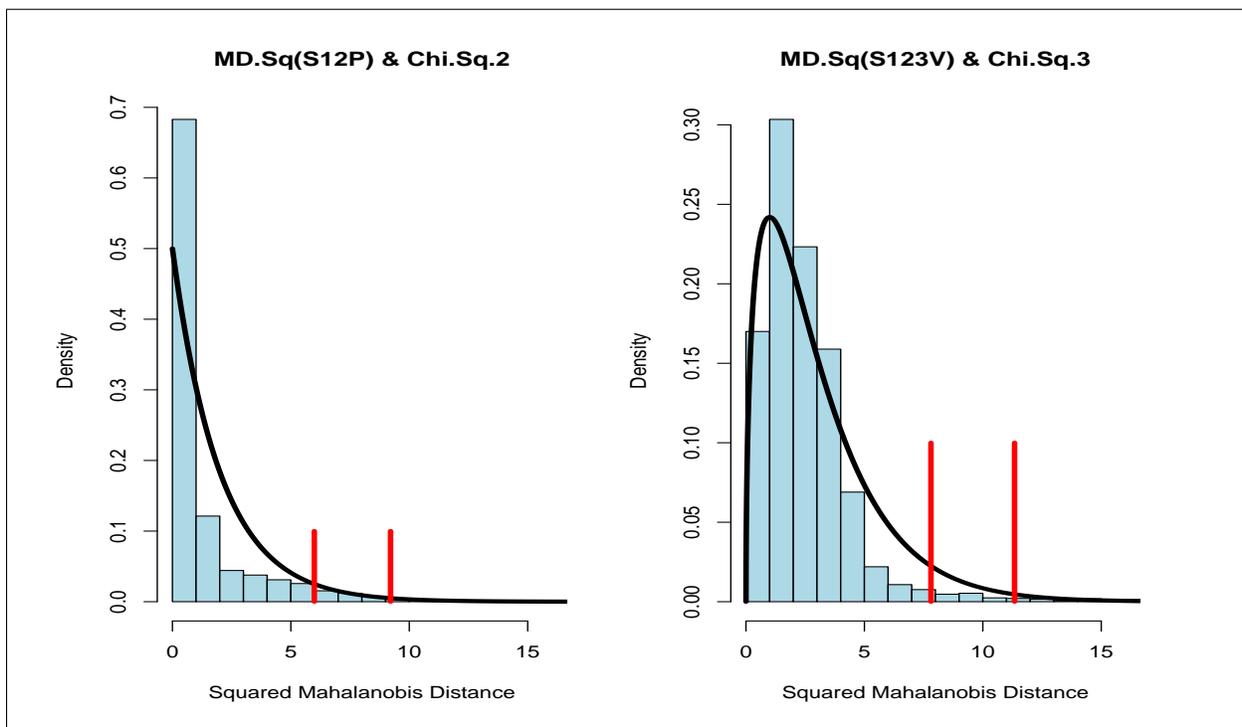
Figure 25: Left panel: Squared Mahalanobis distances of South Lanes 1 and 2 hourly percent-volumes. Right Panel: d_m^2 of South Lanes 1, 2 and 3 hourly volumes. If d_m is multivariate normal, then the distribution of d_m^2 should follow a χ^2 of appropriate degrees of freedom. Curves are superposed χ^2 distributions. Short vertical lines indicate 95% and 99% critical values assuming critical d_m^2 do equal critical χ^2 values. All areas under histograms and curves are equal to one to enable comparison, but all ranges truncated at 16 since there is no visible distinction from the x-axis beyond 16.

Figure 26 is the counterpart to Figure 25 using the data from the northbound lanes. The results are very similar to those of Figure 25 and the conclusions are the same: The empirical and theoretical distributions are very similar especially for values of squared-Mahalanobis distances and χ^2 above 10. This suggests that, at least for our dataset, squareroots of χ^2 critical values can serve reasonably well as threshold or cut-off values for declaring d_m outliers even when the underlying distributions are not exactly multivariate normal. But just how well?

MD.Sq(N12P) & Chi.Sq.2   **MD.Sq(N123V) & Chi.Sq.3**
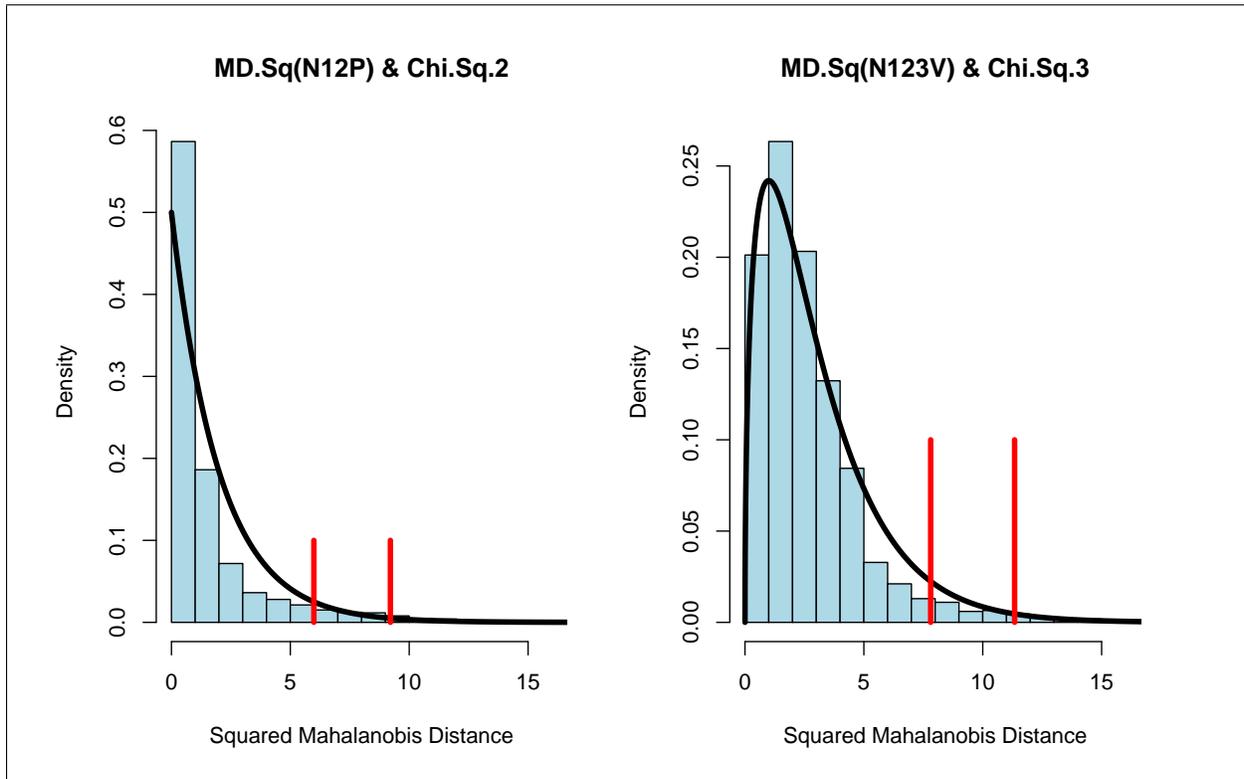
Figure 26: Left panel: Squared Mahalanobis distances of North Lanes 1 and 2 hourly percent-volumes. Right Panel: d_m^2 of North Lanes 1, 2 and 3 hourly volumes. If d_m is multivariate normal, then the distribution of d_m^2 should follow a χ^2 of appropriate degrees of freedom. Curves are superposed χ^2 distributions. Short vertical lines indicate 95% and 99% critical values assuming critical d_m^2 do equal critical χ^2 values. All areas under histograms and curves are equal to one to enable comparison, but all ranges truncated at 16 since there is no visible distinction from the x-axis beyond 16.

5.7 Accuracy of $\sqrt{\chi^2}$ as an Outlier Threshold

Figures 5, 6, 9, and 10 show that, for our data, the assumption of multivariate normality is not strictly tenable. However, Figures 23 to 26 suggest that using the theoretical values of $\sqrt{\chi^2}$ leads to reasonable conclusions. But how reasonable is "reasonable"? More generally, how robust are the results to assumption failure? If there are questions about the accuracy of $\sqrt{\chi^2}$ as an outlier threshold, is there a good alternative procedure?

If, for the d_m values corresponding to the fractional-volumes depicted in Figures 23–26, we use as outlier thresholds the square roots for χ^2 with two degrees of freedom for the .05 and .01 levels, we expect (by definition) to find 5% and 1% of the d_m values to equal or exceed the threshold values. Table 4 lists the χ^2 values, the thresholds, and the expected and observed percentages per traffic direction.

Table 4: Proportion d_m Outliers Using $\sqrt{\chi^2}$ Thresholds: Expected & Per Direction

χ^2_2	$\sqrt{\chi^2}$	Proportion $d_m \geq \sqrt{\chi^2}$		
		Nominal	South	North
5.99	2.45	.05	.057	.070
9.21	3.03	.01	.021	.030

Using the $\sqrt{\chi^2}$ values as thresholds results in declaring as outliers up to 2% more d_m values than would be expected if the underlying distributions were multivariate normal. If you are looking for anomalies, this use of the $\sqrt{\chi^2}$ values as thresholds has proven to be a conservative approach in the sense that possible anomalies have not been overlooked. There is a small cost in that a few "routine" traffic patterns have been flagged for further investigation. In the language of signal detection theory, this procedure has resulted in a high hit-rate at the expense of a moderate amount of false-alarms.

5.8   Distribution-Free Outlier Thresholds

If, instead of starting with theoretical χ^2 values, we start with the actual empirical distribution of d_m values, it is an easy matter to determine the actual d_m values which (on an equal to or greater than basis) yield the highest 5% or 1% (or other percent) of the d_m values. The definition of anomalous value is then determined directly by a researcher and does not depend on any assumptions about underlying theoretical distributions. For example, the d_m thresholds which denote the highest 5% or 1% d_m values of the fractional-volumes in the northbound and southbound directions as anomalous are given in Table 5:

Table 5: Actual d_m Values for Exact Percent Outliers Per Direction

% Desired	Nominal d_m	Empirical d_m	
		South	North
5%	2.45	2.53	2.72
1%	3.03	4.40	3.78

The results in Table 5 are in accord with those in Table 4 but viewed from a different angle. Since the theoretical cut-offs in Table 4 resulted in too-generous a helping of outliers, the empirical cut-offs must be slightly higher to slim down the bulge in declared outliers. All the empirical cut-offs in Table 5 are greater than their theoretical counterparts.

5.9   Detecting Contiguous Multi-Observation Events Using d_m

We have shown that high values of Mahalanobis distance, even when multivariate normality assumptions are not met and there are large numbers of missing data, are useful for identifying individual outliers and anomalous observations.

How to set a threshold for defining "high values" of d_m was covered in the discussion of Tables 4 and 5. In turn, how to use constant d_m ellipsoids with scatterplots of data components for identifying individual-outlier multivariate-observations was covered in the discussion of Figures 23 and 24.

These individual-outlier data-records exhibit anomalous *internal* patterns. Here "internal" refers to the values of the component volumes or fractional-volumes within a one-hour observation. The character of a multivariate data-point being anomalous is independent of the values or internal structure of the other data-points in its immediate spatio-temporal neighborhood. This is the way Mahalanobis distance is typically used for detecting outliers.

But Figures 18 and 19, and more so, Figures 20 and 21, show that there exist contiguous clusters of outlier hours which, viewed as a whole, have an *external* pattern that signals an anomalous event exceeding the span of the units of analysis, i.e., individual Mahalanobis distances. Here "external" refers to the time-series pattern of the Mahalanobis distances associated with the components (and not the components themselves). The character of a contiguous set of Mahalanobis distances being anomalous is heavily dependent on the values of the d_m's in a spatio-temporal neighborhood. In particular, a contiguous cluster of *medium* Mahalanobis distances, none or few of which by themselves would draw attention to themselves, may itself be anomalous depending on its place in a time-series.

The finding of anomalous events which exceed the span of the units of analysis is illustrated by comparing Tables 6 and 7.

Table 6 lists all Mahalanobis distances greater than seven (29 out of 8,282) based on the volumes of the three southbound lanes. The d_m are ordered from highest to lowest with horizontal lines placed between bands primarily indicating basically all 7's or 8's or 9's or reasonable numerical breaks. This ordering by magnitude is reasonable but only shows two contiguous groups of two hours (on April 30 and March 20, 2008). Ordered as the d_m are, it is very difficult to formulate hypotheses about the causes of the outliers such as accidents, repair activity, or weather events.

Table 6: Mahalanobis Distances ≥ 7 for 3 Southbound-Lanes Volumes: Magnitude Order

Date	Hour	S1.v	S2.v	S3.v	S1.p	S2.p	S3.p	MD.S123
30-Apr	1	20	73	1150	0.02	0.06	0.93	12.14
30-Apr	2	0	6	1058	0.00	0.01	0.99	11.98
5-Apr	22	38	45	1076	0.03	0.04	0.93	11.74
12-Apr	22	45	703	1496	0.02	0.31	0.67	11.04
20-Mar	11	262	1543	1462	0.08	0.47	0.45	9.46
21-Nov	0	2	1412	1089	0.00	0.56	0.44	9.41
6-Sep	0	1693	1546	305	0.48	0.44	0.09	9.26
9-Dec	10	811	556	1447	0.29	0.20	0.51	9.19
20-Mar	13	361	1507	1534	0.11	0.44	0.45	8.97
20-Mar	12	321	1474	1474	0.10	0.45	0.45	8.87
5-Apr	23	547	557	1336	0.22	0.23	0.55	8.81
26-Aug	2	1168	29	0.00	0.98	0.02	0.00	8.80
19-Nov	0	0	1280	1071	0.00	0.54	0.46	8.72
18-Mar	11	270	678	1302	0.12	0.30	0.58	8.70
18-Nov	0	3	1276	970	0.00	0.57	0.43	8.39
9-Nov	21	678	654	1393	0.25	0.24	0.51	8.25
18-Mar	12	285	854	1339	0.12	0.34	0.54	8.19
7-Nov	0	1221	1278	72	0.47	0.50	0.03	8.10
12-Apr	23	464	832	1418	0.17	0.31	0.52	8.07
26-Aug	3	989	0	1	1.00	0.00	0.00	8.04
6-Nov	0	1383	1086	0.00	0.56	0.44	0.00	8.03
25-May	16	513	535	1203	0.23	0.24	0.53	8.01
15-Nov	2	1	1204	883	0.00	0.58	0.42	7.82
26-Aug	4	939	2	0	1.00	0.00	0.00	7.76
12-Apr	21	609	1065	1579	0.19	0.33	0.49	7.72
15-Nov	1	244	1361	1125	0.09	0.50	0.41	7.68
18-Mar	13	416	870	1340	0.16	0.33	0.51	7.52
30-Apr	3	1	16	450	0.00	0.03	0.96	7.37
9-Nov	20	774	765	1373	0.27	0.26	0.47	7.06

Table 7 contains the identical data as Table 6, namely, all d_m greater than seven for southbound three lanes volumes, except that the rows are now ordered chronologically. Here single horizontal lines group temporally contiguous hours. Four of these clusters span three hours whereas there were only two two-hour clusters in Table 6. If the outlier threshold had been set lower than seven, some clusters or events greater than three hours would emerge. Some clusters (such as on April 12, 2008) contain observations which carry a relatively high and a relatively low d_m within themselves. It is these groupings that the order-by-magnitude of Table 6 misses.

The value of the chronological groupings is that the clusters both invite and enable suggestions or hypotheses to explain the causes of the anomalous patterns within the groupings. The greater-than-a-single-hour patterns formed by the lane volumes or the fractional-volumes afford a much richer basis for hypothesis generation than just the meager pattern within a single observation. For example,

- On both March 18 and March 20, in the late morning (from 10 a.m. to 1 p.m since the table lists the *end* of the hour-blocks), traffic is maintained but with a large volume and fractional-volume reduction in southbound Lane 1 and the slack taken by an increase in Lane 3. The maintenance of a steady stream in Lane 1 would argue against an accident. Perhaps there was construction or maintenance activity?

- On April 12th, traffic in the late evening is similar to that of March 18 and 20 except that the reduction in Lane 1 is even more pronounced at only 2% hour-relative volume. But, the fact that 45 vehicles still drove past the sensor site argues against an accident. Perhaps there was again some maintenance activity?

- On April 30th, shortly after midnight, traffic is drastically reduced on Lanes 1 and 2 with Lane 3 assuming from 93 to 99% of the traffic. This strongly argues for an accident. We can further surmise that the accident occurred shortly after midnight since 20 vehicles did flow past the sensor site and not all 20 would have been emergency responders. The clean-up probably went into a fourth hour since only one vehicle passed in Lane 1 in the third hour. Figure 21 supports this and indicates that the accident+clean-up event did span four hours, but that the d_m for the fourth hour was just below the threshold of 7 used in Table 7.

- On August 26th, in the pre-dawn hours, there is a similar pattern to that of April 30, but now virtually all traffic is in Lane 1. The zero volume in the first and third hours suggests that the hypothesized accident+clean-up event spanned at least five hours. Again, 7 is a very exclusive d_m threshold since it here flags only 29 of 8,828 units as outliers.

Table 7: Mahalanobis Distances ≥ 7 for 3 Southbound-Lanes Volumes: Chronological Order

Date	Hour.24	S1.v	S2.v	S3.v	S1.p	S2.p	S3.p	MD.S123
18-Mar	11	270	678	1302	0.12	0.30	0.58	8.70
18-Mar	12	285	854	1339	0.12	0.34	0.54	8.19
18-Mar	13	416	870	1340	0.16	0.33	0.51	7.52
20-Mar	11	262	1543	1462	0.08	0.47	0.45	9.46
20-Mar	12	321	1474	1474	0.10	0.45	0.45	8.87
20-Mar	13	361	1507	1534	0.11	0.44	0.45	8.97
5-Apr	22	38	45	1076	0.03	0.04	0.93	11.74
5-Apr	23	547	557	1336	0.22	0.23	0.55	8.81
12-Apr	21	609	1065	1579	0.19	0.33	0.49	7.72
12-Apr	22	45	703	1496	0.02	0.31	0.67	11.04
12-Apr	23	464	832	1418	0.17	0.31	0.52	8.07
30-Apr	1	20	73	1150	0.02	0.06	0.93	12.14
30-Apr	2	0	6	1058	0.00	0.01	0.99	11.98
30-Apr	3	1	16	450	0.00	0.03	0.96	7.37
25-May	16	513	535	1203	0.23	0.24	0.53	8.01
26-Aug	2	1168	29	0	0.98	0.02	0.00	8.80
26-Aug	3	989	0	1	1.00	0.00	0.00	8.04
26-Aug	4	939	2	0	1.00	0.00	0.00	7.76
6-Sep	0	1693	1546	305	0.48	0.44	0.09	9.26
6-Nov	0	1383	1086	0	0.56	0.44	0.00	8.03
7-Nov	0	1221	1278	72	0.47	0.50	0.03	8.10
9-Nov	20	774	765	1373	0.27	0.26	0.47	7.06
9-Nov	21	678	654	1393	0.25	0.24	0.51	8.25
15-Nov	1	244	1361	1125	0.09	0.50	0.41	7.68
15-Nov	2	1	1204	883	0.00	0.58	0.42	7.82
18-Nov	0	3	1276	970	0.00	0.57	0.43	8.39
19-Nov	0	0	1280	1071	0.00	0.54	0.46	8.72
21-Nov	0	2	1412	1089	0.00	0.56	0.44	9.41
9-Dec	10	811	556	1447	0.29	0.20	0.51	9.19

Not all emergent multi-hour patterns have to be strictly hour-hour contiguous. For example,

- The double horizontal lines near the bottom of Table 7 set off three hours one-day apart and each ending at midnight on consecutive days (November 18–21). All show an almost total absence of traffic flow in Lane 1 accompanied by high absolute volumes in the outer two lanes. That the middle lane carried more traffic than Lane 3 suggests that the cause of the lack of traffic in Lane 1 was benign since some traffic would shy away from an accident.

- Returning to March 18 and 19, we may discern a "cluster of clusters" meta-pattern. As argued just above, the intra-cluster three-hour patterns suggest a non-accident construction activity. That the inter-cluster pattern takes place at exactly the same time argues for a planned traffic-slowing, but not traffic-eliminating, activity in Lane 1. The absence of March 19 weakens the argument, but could be explained by either too-high a d_m threshold here or a holiday.

One other class of traffic-reduction events not apparent in Tables 6 or 7 but which should be noted are weather related events. Snow, of course, can reduce traffic across *all* six lanes in both directions for several hours or longer. At a different season, a severe thunderstorm with blinding rain can reduce traffic across all six lanes by slowing it down for, say five or 10 minutes once or twice within an hour. Both of these situations might be detectable by computing Mahalanobis distances using traffic on all six lanes simultaneously as the input variables.

These speculations are hypotheses and as such are, of course, subject to verification using external records. But without first formulating the hypotheses, it would make little sense to seek their verification.

6   DISCUSSION & CONCLUSIONS

It is well-known that Mahalanobis distance is a useful way to transform a set of n multivariate-normal observations into a set of n univariate "characterizations" of the observations. One characterization is that d_m is a standardized measure of how far a multivariate point is from the centroid of an ellipsoid scatterplot such that all points with the same d_m are equiprobable no matter in what direction the point is from the centroid. Another characterization is that high values of d_m arise from multivariate-normal observations having an internal pattern unlike the patterns found in the bulk of the members of the dataset. It is also well-known that, under the assumption of multivariate normality, the *square* of the Mahalanobis distance follows a χ^2 distribution, and thus can be used to detect outliers in multi-variate normal data.

6.1   Principal Conclusion

The principal conclusion of our analysis is that high values of Mahalanobis distance are still useful for identifying individual outliers and anomalous observations even when multivariate normality assumptions are not met and there are large numbers of missing data.

Further, if the source data and observations are a time series, Mahalanobis distance can be used to identify higher-order meaningful yet anomalous groupings even when the component observations are not, taken individually, overly deviant in themselves. This is interesting because the techniques let us find interesting structures and events that exceed the span of the individual units of the analysis. Indeed, the technique of preserving chronological rather than magnitude order when examining a Mahalanobis-transformed time series (such as in Figures 20 and 21 or as in Table 7) can reveal even higher-order structures such as clusters of clusters.

6.2   Contextual Sensitivity of Traffic Volume Data

Another feature of our study is the use of vehicular traffic-volume data from a major highway in a major urban area. Volume data is especially interesting since high or low levels, in themselves, do not have a fixed or unambiguous meaning either in evaluative (e.g., good, bad) or interpretive (e.g., weather event, accident) terms. Rather, the significance of "high" or "low" volume depends on the temporal pattern-of-life context. As Figure 7 shows, there is a complex diurnal fluctuation in traffic volume with low volume in the wee hours and higher volume in the daytime. The daytime hours themselves show a variation with sub-peaks for the morning and evening rush hours. Hence, the definition of "anomalous" volume depends on the temporal context. For example, a given volume level might be abnormally low at rush hour but abnormally high before dawn. A further complexity is that what constitutes anomalous volume also depends on the volumes in the other lanes in the same and in the opposite direction.

The contextual significance of a volume level becomes even more complicated when other cyclical and non-cyclical factors enter into the causes. Weekends, holidays, major sports events, construction, weather, and accidents all affect volume, sometimes increasing and other times increasing volume from typical hourly baselines.

Whether weather, accidents, and road repairs are truly anomalous or normal parts of traffic routines since they do occur with some historical frequencies is subject to context and purpose. We choose to define these events as anomalous since their exact time of occurrence is not always predictable and their impact, from a driver's point of view, is disruptive and not what drivers' want to be normal and routine. Furthermore, high Mahalanobis distances are associated with these types of events, whereas high values of d_m are absent when these events are absent.

6.3   Even Larger Contexts: Multiple Roads & Sensor Sites

Another contextual factor is the complex and busy road network in which I–95 is located located. Accidents, repairs, and sheer traffic volume on neighboring roadways and surface streets can have an impact on traffic volume passing through the one single sensor site whose data we have been analyzing. We have ignored the problem of greater spatial outliers arising from a larger road network context, but for a treatment of such larger-context outliers see Shekhar, Lu, and Zhang (2003) who used multiple-site multiple-road traffic data from another major urban area, namely, Minneapolis.

6.4   Hypothesis Verification & Ground Truth Data

We have advanced several general hypotheses about the Mahalanobis distance signatures for several classes of outlier events affecting traffic volume. We have even ventured specific causes to explain specific d_m patterns on some specific dates and hours. But we have not attempted to verify the hypotheses using external ancillary public records such as accident and weather data. We have not even linked any of the dates to the calendar for 2008 to identify work days, weekends, and major holidays.

This is deliberate. We wanted to see what the volume data, by itself, could say or suggest about its origins. That is, to what degree is the internal structure of a large ($n > 8,000$) and structurally rich (six lanes and two directions) but flawed (large amounts of missing stat) and dimensionally-limited (only volume data, no speed data) dataset sufficient to reverse-engineer the data?

6.5   Reverse Engineering Using Only the Data's Internal Structure

The impact of temporal and non-temporal factors on the meaning of volume data is what makes "reverse-engineering" the volume records so challenging and interesting.

One purpose of this report and analysis has been, in a sense, to develop a technique, based on Mahalanobis distance, to reverse engineer the volume data to determine the influences on the data. We have shown how to nominate anomalous events and hypothesize about their nature such as weather, or construction, or accident events.

These classes of events presumably have signatures, albeit not necessarily unique, such as those discussed in the previous section. Some signatures may be written bold but still not be unique since, for example, zero-volume on all lanes and all directions can be due to either snow or a major truck jack-knifing. The duration of the event might help disambiguation but that invokes the greater-than-basic-unit-of-analysis contiguous cluster technique we discussed

earlier. Other signatures may be subtle, faint, and indirect as in our speculation about evidence for rubbernecking from the possible discernment of volume reduction in the nearest lane in the opposite direction to where there was (or might have been) an accident.

6.6   Future Work

This is an exploratory report concerned with hypothesis generation and prediction. Future work will focus on questions of assessment, validation, and generalizability. Some questions to be explored are:

- **Ground truth.** A follow-on report will use external, public records of weather, accidents, holidays, and special events to provide ground-truth and testing of the proposed explanations of the volume-only data to account for itself.

- **Quantification of accuracy.** Although we anticipate some success, it will not be perfect. We do want to quantitatively measure the ability of Mahalanobis distance to accurately identify truly anomalous events. By accurately identify, we mean correct identifications (hits), missed events (misses), correctly-called uneventful events (correct rejection), and wrongly-called events (false alarms). The language of hits, false alarms, correct rejections, and misses immediately permits us to use the quantitative metrics of detection ability and bias offered by the Theory of Signal Detection.

  The signal detection metrics can then let us assess and rank-order the value of various data-streams (such as speed data) which may be added to better identify and predict outliers and anomalous patterns in event data.

- **Performance with sparse data.** We have shown the value of Mahalanobis distance with large datasets to flag outliers and outlier contiguous-clusters even when its underlying normality assumptions are not met. But there remain questions to be explored such as assessing its accuracy and tolerance using sparse and very sparse datasets. This may be done with the existing dataset by systematically introducing gaps and truncations. The outlier detection performance of the reduced datasets may then be compared with the performance of the current dataset. How much data needs to be dropped to reduce performance by $x\%$? Performance for various values of x may then be graphed.

- **Stability of the covariance matrix.** The question of performance with sparse data is related to the stability of the covariance matrices Σ and S which lie at the heart of the definition and computation of Mahalanobis distance (see Eqs 1 and 2). Since the covariance matrices used in this report are all based on over 8,000 data points, they would be expected to be fairly stable even with the change, addition or deletion of, say, 10% of the data-points. The question is empirical and there are statistical tests for comparing covariance matrices.

- **Refinement of the covariance matrix.** As discussed in the section on Mahalanobis distance, one of the attractions of the technique is that various robust alternatives may be used for the covariance matrices in Eqs. 1 and 2. We earlier described one

procedure in which the data point whose d_m is being computed is removed prior to computing the covariance matrix so as not to have the effect of an outlier pollute or degrade the matrix and thus skew the results. With a huge n, such as in this report, it would be possible to use the full data to identify, say, 0.5% of the most extreme d_m and then compute a second covariance matrix with the extreme points removed. This second-generation covariance matrix can then be used for the analysis. Again, the equality of the first- and second-generation covariance matrices may be evaluated.

- **Ability to extrapolate.** Extrapolation is something to be approached carefully. But the reason behind considering questions of sparseness, covariance matrix stability, and covariance matrix refinement is to be able to predict with confidence. Assuming some general stability in the covariance matrices based on one year's data, it should be possible to extrapolate into the future unless there is reason to believe in some change in driver behavior or change in traffic laws and conditions.

- **Detecting subtle effects.** The final question is about the detection ability of the Mahalanobis distance technique. Almost any detection technique should be able to spot a multi-lane multi-direction disaster shutting down an Interstate highway for hours. But the real test of a technique is its ability to detect subtle, yet atypical effects. For example the switch to Daylight Savings Time and back affects the sleep patterns and daylight expectations of drivers and is known to, in turn, affect driving (Vanderbilt, 2008). Can Mahalanobis distance detect any change in traffic patterns on the first morning rush hour after a time shift? On later days?

A final note: Mahalanobis distance is only one on many outlier detection techniques. Any serious analysis of traffic or other patterns should, of course, use more than one technique.

REFERENCES

Barnett, V., & Lewis, T. (1994). *Outliers in statistical data* (3rd ed.). New York: Wiley.

Beaty, W.J. (2011). *Traffic waves.* Retrieved March 7, 2011 from http://trafficwaves.org

Camazine, S., Franks, N.R., Sneyd, J., Bonabeau, E., & Deneubourg, J.-L. (2003). *Self-Organization in biological systems.* Princeton, NJ: Princeton University Press.

Flury, B. (1997). *A first course in multivariate statistics.* New York: Spinger-Verlag.

Johnson, R.A., & Wichern, D.W. (Eds.). (2007). *Applied multivariate statistical analysis* (2nd ed.). Upper Saddle River, New Jersey: Pearson Prentice Hall.

Mahalanobis, P.C. (1936). On the generalized distance in statistics. *Proceedings of he National institute of Sciences of India, 2,* 49–55.

Miller, J.H., & Page, S.E. (2007). *Complex adaptive systems: An introduction to computational models of social life.* Princeton, NJ: Princeton University Press.

NYSDOT (2011). "Hourly Traffic Data." New York State Department of Transportation, Engineering Division, Office of Technical Services. URL: https://www.nysdot.gov/divisions/engineering/technical-services/highway-data-services/hdsb Accessed 3 Jan 2011.

Rousseeuw, P.J., & Leroy, A.M. (1987). *Robust regression and outlier detection.* New York: Wiley.

Shekhar, S., Lu, C-T., & Zhang, P. (2003). A unified approach to detecting spatial outliers. *GeoInformatica, 7,* 139–166.

Vanderbilt, T. (2008). *Traffic: Why we drive the way we do (and what it says about us).* New York: Alfred A. Knopf.

LIST OF ACRONYMS

DOT	Department of Transportation
I–95	Interstate Highway 95
LS	Least Squares
MB	Mega Byte
MD	Mahalanobis Distance
N1	Northbound Lane 1
N2	Northbound Lane 2
N3	Northbound Lane 3
NYSDOT	New York State Department of Transportation
Q–Q	Quantile-Quantile
RC_ID	Road and County Identifier
S1	Southbound Lane 1
S2	Southbound Lane 2
S3	Southbound Lane 3