Final Report

Cellular Probe Data Evaluation

Case Study: The Baltimore Multimodal Traveler Information System (MMTIS)

Ali Haghani

Saini Yang

Masoud Hamedi

January 2007

University of Maryland

Executive Summary

This work was performed as part of the Multimodal Traveler Information System (MMTIS) project. This project is a partnership between Maryland Department of Transportation (MDOT) and a private sector team led by Delcan Corporation.

This report summarizes the findings of an independent study sponsored by MDOT, Delcan Corporation, Federal Highway Administration (FHWA) and the I-95 Corridor Coalition. The contents of this report reflect the views of the authors. This report does not constitute a standard, specification, or regulation.

Development of wireless technology and the increasing penetration rate of cellular phones provide a potentially enormously valuable tool to improve the performance of transportation system. When the drivers' locations can be transformed to real time trajectories and mapped to road network, travel speeds and travel times can be estimated and provided to real time traffic control or traffic management centers and used in traveler information systems. Cellular probes can provide complementary information for freeway networks and major arterials as well. Currently, only 35% of urban freeways have some form of surveillance and cellular probe information brings a much higher coverage for the entire freeway networks. For those freeways that have single loops that are not capable of producing very accurate traffic speed and density measurements, cellular probe information can provide more accurate measurements. Cellular Probe data can also be used for rural highways, minor urban highways, local arterials and streets, where no instrumentation is planned.

The purpose of the study is to investigate the quality of data that is obtained using cellular technologies in generating real time traffic information for use in regional traffic control and management and traveler information systems. Baltimore metropolitan area is the focus of the case study. Two major data sources were used in evaluation: traffic information based on cellular phones probe data provided by ITIS Holdings and probe vehicle data provided by MotionMaps LLC. The Coordinated Freeways Action Response Team (CHART) sensor data downloaded from the website of the Center for Advanced Transportation Technology Laboratory (CATT LAB) at the University of Maryland, College Park and was used for data reference purposes. The evaluation results presented in this report are based on the available data.

Cellular probe traffic data provided ideal spatial and temporal coverage for the entire Baltimore metropolitan road network during the test period of 21 days. This feature enables great potential for future ITS applications, such as traveler information system and real-time traffic control. None of the other traffic data collection mechanisms provide a coverage that is comparable to the cellular probe technology at this time. The cellular probe data provided traffic speed and travel time information every 5 minutes. The accuracy of the travel time data obtained from cellular probes on freeways is quite good (around 10% error terms) for immediate ITS application. The accuracy for travel speed data obtained from the cellular probes on freeways is acceptable (around 20% error terms). On arterials, the cellular probe data cannot provide accurate travel speed and travel time information at this time. However, this may have been influenced by several factors that will be discussed later. Future improvements of technology and incorporation of other sources of data (such as GPS data) can further enhance the accuracy of data for ITS applications.

The study encompassed designing a survey plan that can provide a systematic evaluation base, investigating how well the cellular probe data represents the real traffic situation, and comparing the cellular probe data with the sensor traffic data. In order to accomplish its goals, the study consisted of three phases: Phase I – survey design and initial field test, Phase II – survey data collection and processing, and Phase III -- statistical analyses.

The evaluation methodology and the statistical tests were carefully selected upon the nature of the data. Since in most cases the normality assumption required for standard statistics tests is not satisfied for the variables of interest, various nonparametric tests were explored in this study. In the designed survey plan, the survey routes cover freeways and arterials, the survey areas include suburban and downtown CBD areas, and the survey times cover weekdays, weekends, morning peaks, afternoon peaks and non-peak hours. A total of 500 probe vehicle survey runs covering 9,000 miles were conducted during 26 January to 3 February 2006. Due to budget limitations 2 to 3 probe vehicles were used for a survey route in a certain time interval. The TMC[1] (Traffic Message Channel) road network was used and the trajectories of probe vehicles were matched to the TMC nodes for data quality evaluation. The cellular probe data of Baltimore metropolitan area spans from 26 January to 16 February 2006. The travel speed and travel time on each link are provided in every 5-minute interval. The vehicular probe data of each run is provided as a raw GPS file. MotionMaps also matched the coordinates of probe vehicle locations to the TMC node number and provided the data in database files.

It is important to point out that there are several issues regarding to the data quality evaluation. These issues can bring large variability to the data and greatly degrade data quality measures. Clarifications of these issues can help readers to understand the data reported in this evaluation project.

(1) The TMC road network refers to a set of roadway link that is used in commercial navigation systems.

"Ground Truth": in this study, the probe vehicular data is treated as the "Ground Truth" in the calculation of data quality measures. In reality, there is no universally accepted measure of "Ground Truth" and each measure has advantages and disadvantages. Theoretically, probe vehicles can provide average travel speed when drivers strictly follow floating car method. In practice, when the traffic is congested on road, it is difficult for a driver to pass one car for every car that passes him or her. Another problem is that when there are many lanes on a freeway segment, it is difficult for the driver to observe the movements of all vehicles in every lane. Similar problems occur when traffic signals are installed on streets since vehicle speeds are affected by whether or not their travel is coordinated with traffic signals. Statistical sampling methods suggest that roughly 4% to 5% of probe vehicle penetration can provide a good estimate of travel times based on simulation results [1]. In our study, the number of probe vehicles is far less than the recommended values due to budget limitations. Therefore the vehicle probe data that are used in this study may not accurately represent the "Ground Truth".

As it is quite possible to match a single segment with only 5 or 10 data points, the conclusions here are drawn on relevance and similarity rather than correctness of traffic information on road segments. However, the values of data quality measures reported are still based on the "Ground Truth" assumption.

Default Values: according to ITIS Holdings, on some routes where there is no real time data available, default value (usually, the speed limit) are provided as cellular probe speed. These default values sometimes are far from the vehicular speed and thus generate large errors and degrade the data quality measures. The deficiency of cellular data may be caused by low penetration of cellular phone or limitations of cellular carrier. This issue might be solved by contracting with a cellular carrier with high market penetration rate.

Outlier Impact: this impact is caused by the extremely slow speeds. The vehicular data is used as the "Ground Truth" in the evaluation. When calculating the percentage of errors, probe vehicle speeds are used in the denominator. Huge percentage of errors will be generated when the absolute error between the cellular speed and vehicular speed is relatively large. As an example, when the vehicular speed is 5 miles per hour and the cellular speed is 20 mile per hour, the absolute error is 300 percent! These outlier data points contribute to increasing the average percentage of absolute error. Therefore, it is important to pay attention to the histogram analysis of the data when drawing conclusions from the percentage of average and absolute errors that are presented in this report.

The evaluation of travel speeds was performed on TMC links (Traffic Message Chanel), which are the most detailed roadway links used by commercial customers and the smallest unit in cellular probe data, because the link discrepancies can have major impact on the link travel time comparisons. The probe vehicle speeds are deemed "Ground

Truth" when calculating the data quality measures. The average absolute errors on freeways are about 10 miles per hour (8.0 miles per hour on I-70, 9.8 miles per hour on I-95, 12.4 miles per hour on I-395, 8.1 miles per hour on I-695 and 10.3 miles per hour on I-895). The median of absolute errors on freeways are in a range of 5.8 miles per hour to 9.7 miles per hour. The statistical tests results show no significant difference between cellular link speeds and vehicular link speeds on all freeway routes, except I-70. The fact that the cell probe data overestimated speeds relative to the drive tests is not surprising since ITIS used speed limits as a default value. During peak travel times, the speed limit usually exceeds the actual travel speeds. It should be noted that the relatively large average error measures do not necessarily mean dissimilarity in data. Some extremely large absolute errors can dramatically degrade the data quality measures of a data set. It is important to refer to the corresponding histogram when reading the values of data quality measures in order to better understand and evaluate the data quality.

On arterials, on the other hand, the results of statistical tests on link speeds show that there are significant differences between cellular probe speeds and vehicular probe speeds. The travel speed obtained from cellular probe data is significantly larger than those from probe vehicle data. The average vehicular speeds on arterials are around 20 to 25 miles per hour. The cellular probe speeds consistently overestimate vehicular link speeds. The overestimations range from 17.4 to 24.4 miles per hour, which can be more than 100% of the average vehicular speeds on some routes. The median of absolute errors ranges from 14.8 miles per hour to 24.9 miles per hour. As an example, on MD Route 40 which runs through the CBD area in Baltimore, the average absolute error is 17.4 miles per hour, which is more than 60% of the average probe vehicle travel speed. The statistical tests results confirm the dissimilarity of these two types of speeds on all arterials. The assumptions discussed above may account for part of the difference.

For paths that are composed of multiple links travel time evaluation was conducted. Travel time evaluation showed the same trend in link travel speed evaluation. That is, cellular probe data provides good estimations of path travel times on freeways. The deviations between path travel times obtained from cellular probe data and those obtained from vehicular probe data are about 10%-- not that far off from the variations found when comparing individual test runs. Therefore, the cellular probe travel time information can be directly applied in traveler information systems. The path travel times obtained from the cellular probe data is significantly larger than those obtained from probe vehicle data on arterials. On MD Route 40, the average deviation is more than 40% of the probe vehicle travel speeds.

The comparison among cellular probe data, sensor data and vehicular probe data are based upon traffic speeds. Due to the data availability, no travel time information is available from sensor data. The statistical test results show that in most cases all three technologies produced individual data points that were different from the others. The cellular data outperforms the sensor data, especially in capturing the overall picture of the variations in vehicular speeds.

Cellular phone tracking, at least the one implementation we examined, shows promise and would benefit from additional technology improvement. It also needs further study to determine its ability to produce quantifiable travel time data. On the positive side, it promises very large numbers of probes and better spatial and temporal coverage than conventional loop detectors. However, the evaluation results indicate that there is a direct relationship between the road type and quality of estimated traffic information. On most freeways, the evaluation results are consistently good. On arterials on which traffic signals exist, the cellular probe data tends to over-predict the travel speeds and under-predict the travel times significantly. Another issue, which is particularly problematic for cellular probe data, is the congestion and data quality relationship. When the level of service is low, the differences between the travel speeds and travel times obtained from the cellular probe data are larger deviation than those obtained from the probe vehicle data compared to when the level of service is high. The ITIS data may perform better when adequate data exists, as shown by most freeway results. In addition to improving the underlying algorithms, other avenues to correct this problem include: 1) not using speed limits as default values for speed when data is not available; 2) working with a wireless carrier that can provide greater market penetration; and 3) using additional available data , for example GPS data from local fleets, in estimating the speeds.

For cellular phone tracking systems further examination of the nature of errors, particularly additional analysis and comparison of data for arterials, is needed before any error correction algorithms could be developed. Additional information such as signal phase setting could be useful and should be provided for analyzing cellular probe data quality.

Table of Content

List of Tables

List of Figures

1 Introduction

1.1 Background

Various data collection techniques such as loop detectors, Closed Circuit TVs (CCTV), cellular phones and GPS units, can be used to obtain real time traffic information. While major freeways in the Baltimore metropolitan area are equipped with inductive loop detectors and CCTV cameras, there is no plan to instrument minor or rural freeways. Many municipal streets in the downtown area do not have the infrastructure needed to monitor arterial traffic flow. Furthermore, nationally on the average typically about 50% of loop detectors are producing accurate or reliable traffic data due to a variety of reasons including hardware, software and communication problems [1].

Alternative methods for traffic data collection are needed for two important reasons: 1) improving the surveillance data for traffic management and operations and 2) improving the dissemination of accurate and reliable traveler information.

By the end of 2005, the total number cellular phone user in the U.S. was over 207 million [2]. This means the penetration rate of cellular phone is more than 69 percent of the U.S. population. The proliferation of wireless technology is providing Intelligent Transportation Systems (ITS) operators with an enormously valuable tool to improve its performance. Certain technology has been developed to determine cell phone carriers' real time trajectories and map them to vehicles' movements on road networks.

Cellular probe information can complement the information collected from single loops on freeways that are equipped with such single loops since the single loops are not capable of producing very accurate traffic speed and density. Cellular probe data can also be used for rural freeways or minor urban freeways where no instrumentation is planned. Local streets and arterials are also good candidates for collecting data from probe vehicles. Considering the slow progress of loop detector instrumentation of traffic surveillance systems and the extent to which loop systems require maintenance, cellular phone probe systems can be considered as viable means of supporting traffic data coverage, especially in those areas where the existing detectors do not work properly and where instrumentation may not be deployed in the near future. The cellular probe systems have the potential to provide cost-effective wide-area data coverage including both freeways and arterials. No heavy infrastructure investment is necessary for cellular probes.

In summary, when compared to conventional traffic data collecting technologies, cellular probe technology has numerous benefits:

- Immediate wide-spread coverage of the entire roadway.
- Elimination or drastic reduction of the need for fixed detection surveillance investments.
- Lower capital and maintenance costs than traditional sensors.
- Capability of providing link travel times for real time travel time estimation.
- Data availability regardless of road surface and weather conditions.
- Ease of integration of the collected data with existing data from other surveillance technologies.

The reliability and precision of cellular probe data are the most important issues for application of this technology. Since traffic information derived from cellular probe data are typically used in traffic control and management or provision of traveler information system, the precision of the cellular probe data impacts the reliability of real time traffic information and speedy emergency assistance to those in need. Beside the potentially enormous advantage of cellular probe data, there are some concerns with respect to the cellular data quality.

Sampling concerns: Statistically, the ideal case is to investigate the entire population using the same methodology. Sampling error is the difference between an estimate based on a sample and the corresponding value of the entire population. By using cellular data, it is assumed that the cellular signals being processed can represent the integrity of entire population. Sampling error in data may arise when there is demographic characteristics bias of cellular phone users, or when the number of subscriber of one carrier is too small to represents the entire population of drivers.

Positioning concerns: Generally, the more accurate the cellular phone positioning, the less the potential for ambiguity in the probe data. The accuracy required for an effective probe system is dependent on the geography of the road network. In areas with roads are far apart with few intersections, a much lower degree of accuracy in position identification can be used that still can yield correct road segment matches; while in areas with higher density, with roads close together and many intersections, much greater accuracy is required for the probe to be correctly placed on its road segment. In CBD area, the interference of pedestrians who carry cellular phone may cause problems for matching signals as well.

Other concerns: The signal quality may influence the probe data quality. Another major concern with cell data is that some parts of the country have no cell service. Similarly, when the signals are blocked by geographic obstacles along roadside, or are impeded by bad weather, the cellular probe data quality may be affected.

Therefore, it is necessary and crucial to verify the quality of cellular probe data before applying the technology in large scale real world operation.

1.2 Project Objectives and Scope

In order to improve the state of the art and practice in traffic control and management, this new traffic monitoring method and tool needs to meet the challenge of delivering quality data. The objective of this project is to provide a systematic evaluation of the quality of the cellular probe data, as compared with other reliable sources, and mainly with vehicular probe data.

To achieve this objective, a comprehensive analysis is necessary to develop a study framework for data quality evaluation; to identify the factors that potentially influence the probe data quality; to select a reasonable sampling methodology and sampling scheme which take all those potential factors into account while reducing the workload of survey; and to select appropriate statistical models to compare and evaluate different data sources. The focus of this project is to compare the quality of the traffic data produced by the cellular probe technology that is implemented in the case study of Baltimore area with the data collected by vehicular probes. The concepts of data quality measurement are sufficiently developed and adequate data exist to allow such a comparison to take place. Therefore, most of the effort in this project is devoted to the case study. As such, the report includes the following essential elements:

- Identification of factors that may affect cellular probe traffic data quality in Baltimore metropolitan area;
- Selection of data quality measures and statistical analyses;
- Demonstration of source data properties;
- Evaluation of cellular probe traffic data quality.

1.3 Organization of Report

The remainder of the report is organized as follows: Chapter 2 presents the research approach and methodology. Chapter 3 presents the summary of the data used in this study. This chapter illustrates the availability and format of cellular probe data, vehicular probe data and sensor traffic data. Chapter 4 presents the results of the evaluation of data quality on links and paths. Cellular travel speeds on links and travel times on paths are evaluated based on a comparison with vehicular probe data. The data presented in this report is based on the sample and are subject to sampling variability. Chapter 5 presents the results of comparison of travel speeds obtained from sensor data, cellular data and vehicular data. Finally, Chapter 6 presents the concluding remarks.

2 Methodology

2.1 Data Quality Measures

"Data quality is the fitness of data for all purposes that require it. Measuring data quality requires an understanding of all intended purposes for that data [3]."

In this project the cellular probe data is to provide real time system-wide travel time and traffic speed. In 1999, ITS America and the U.S. DOT convened numerous stakeholders meetings and developed guidelines for quality advanced traveler information system data [4]. These guidelines supported the expansion of traveler information products and services, explicitly, to control the quality of traffic data being collected. Among the seven data attributes, six were selected as data quality measures in the report produced by Office of Policy, Federal Freeway Administration in 2003 [3]. These include:

- Accuracy – how closely does the data collected match actual conditions?
- Confidence – Is the data trustworthy?
- Delay – How quickly is the data collected available for use in ATIS applications?
- Availability – How much of the data designed to be collected is made available?
- Breadth of Coverage – Over what roadways or portions of roadways are data being collected?
- Depth of Coverage (Density): How close together/far apart are the traffic sensors?

In this project, we are not able to measure the delay since the cellular data is provided by ITIS Holding as real time information. As no traffic sensor is involved in cellular probe technology, the depth of coverage measure is not necessary. Hence, the accuracy, confidence, availability and coverage are selected as major data quality measures.

Among these measures, Tarnoff [5] suggests data quality measures and possible requirements on speed accuracy and availability (as shown in Table 2-1).

2.2 Existing Similar Projects

Yim, et al. (1991) [6] discussed the potential of using probe vehicles and loop detector data as base scenarios in cellular probe data quality evaluation. In 2001, STL [7] reported a moderate quality of link travel speeds by wireless location technology (WLT). Point sensors were used as the baseline measurement approach in their evaluation. A van-mounted video detection system was used for baseline data collection. The video detection system was used to derive point measures (spot speeds and counts) by processing video from a camera mounted on a 45-foot telescoping mast installed on the van.

Table 2-1
Possible INFOstructure performance requirements

Measure	Application	Requirement	
		Local Implementation	National Implementation
Speed Accuracy	Traffic Management	5-10%	5-10%
	Traveler Information	20%	20%
Volume Accuracy	Traffic Management	10%	N/a
	Traveler Information	N/a	N/a
Timeliness	All	Delay < 1 minute	Delay < 5 minutes
Availability	All	99.9% (approx. 10 hours per year)	99% (approx. 100 hours per year)

Several companies are generally considered to be major market sharers in cellular probe technologies. These include AirSage, Applied Generics, Cellint, IntelliOne and ITIS Holdings. Table 2-2 shows the location of deployments of cell phone probe systems in the U.S., the company involved, and the evaluation status, if known by Spring, 2006 [8].

Due to the timeliness of these projects, no report or working paper is available to describe the methodologies used in these cellular probe data evaluations, except the interim report from ITIS Holdings for Missouri DOT [9].

Table 2-2
Existing deployment of cellular probe systems in US

Location	Company	Evaluation Status
Norfolk, VA	AirSage	Evaluation completed by Virginia Transportation Research Council
Atlanta, GA	AirSage, Cellint	Unknown
Tampa, FL	IntelliOne	Unknown
Baltimore, MD	ITIS Holdings	University of Maryland
Missouri	ITIS Holdings	Phase I of deployment underway
Florida* (study only)	Various study underway at Florida International University	Comparison of various companies' data to existing sensor data underway

2.3 Evaluation Framework

In this evaluation, vehicles equipped with GPS systems are used to collect traffic information in real time while traveling in the traffic stream. Theoretically, when drivers strictly follow floating car method, the average travel speed on a road segment should be obtained. In practice, when the traffic is congested on road, it is difficult for a driver to pass one car for every car that posses him or her. Another problem is that when there are many lanes on a freeway segment, it is difficult for the driver to observe the movements of all vehicles in every lane. Similar problems occur when traffic lights are installed on streets. These factors should be identified and the generated travel times should be treated with extra care or the accuracy of the evaluation will be degraded. All these difficulties can be partially compensated for by increasingly sophisticated post-processing of the data stream which can increase the accuracy of the results by discarding problematic data. The number of probes should be increased in order to compensate for the lost data or bad data quality.

Beside probe vehicle data, loop detector/sensor data are used in this data quality evaluation. Due to detector health and hardware reasons, not all detectors/sensors can provide accurate and reliable traffic information. The detector/sensor data is used for data reference purposes rather than as "Ground Truth" of traffic information.

The flow chart of the data evaluation procedure is shown in Figure 2-1. The major tasks in this project included:

- Initial study: in which the preliminary information such as the required data type and data format was collected, and the important factors that may influence the cellular probe data was identified.
- Sampling scheme design: based on the identified factors the sampling scheme was selected.
- Data collection and data control: these two sections were closely related to each other. The data quality control during data collection helped modify the sampling scheme in order to reduce potential waste of efforts.
- Statistical analyses: the collected data was tested by performing statistical analyses. The correlation and consistency of data from different sources was evaluated.

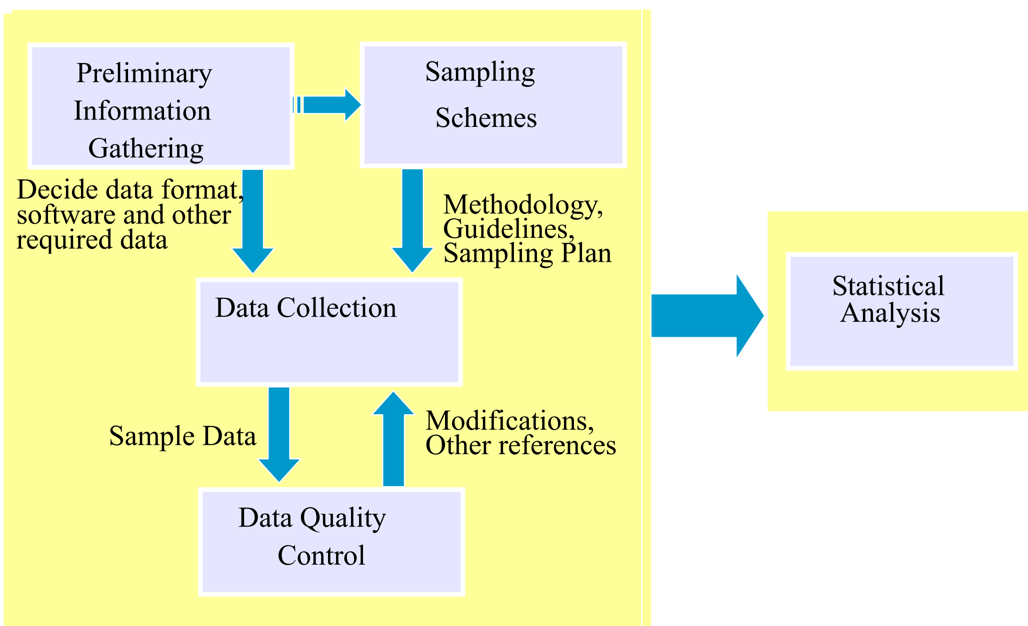

Figure 2-1
Framework of evaluation procedure

2.4 Important Factors

Various variables are considered when identifying significant factors that will influence planning, sampling and evaluation. The following are the ones selected to be tested in evaluation:

- Time of the day/week: the impact of peak hour and non-peak hour, work day and weekend.
- Origin and Destination: the impact of route distance, special network structure.
- Direction: influence of tidal effect of traffic.
- CBD area and pedestrian: the impact of possible disturbance from cell phones carried by pedestrians and bicyclists.
- Weather: the impact of undesirable weather on cellular signal and data quality.

The significance of each factor can be determined by conducting statistical analysis on historical data and significant factors were considered in developing the sampling strategy.

2.5 Sampling Strategy

One important task in this project was to determine when and where the probe vehicle should do survey in the network. As it is generally infeasible to test an entire survey area, sampling strategies must be studied before providing survey scheme.

Sampling strategies can be classified as either non-probabilistic or probabilistic [10]. Non-probabilistic sampling is used in areas in which one is most interested or in already known or suspected sites and where there is no need to sample elsewhere. Probabilistic sampling is used when it is necessary to have a representative sample of the segment in a network, but it is possible to sample only a small percentage of the whole. If it were possible to examine the entire sample universe, the result will be complete with less random error. Unfortunately, it would be extremely expensive to have the opportunity to engage in such complete testing. Simple random sampling, which is the simplest form of probabilistic sampling, has a great drawback that depending on the dispersion of the randomly selected numbers, large parts of the region may be left out of the sampling completely.

Stratified random sampling is one form of probabilistic sampling; it attempts to minimize variability within different zones (or "strata") in the sample universe. The sample universe is divided into large virtual layers and each is designated the amount of sample units proportional to its importance, e.g. time of day can be one layer and the road type can be another layer. The position of units within each layer is determined by random sampling.

When sample units are evenly distributed throughout the sample universe, it is called systematic sampling. This method can avoid the problem of having areas of low sample concentration.

In this study, each identified factor is deemed as one layer and a systematic sampling strategy is applied when developing the survey scheme. The detailed survey scheme is presented in Chapter 3.

2.6 Statistical Tests

2.6.1 Errors in Data

Before performing statistical tests, it is important to understand the potential errors in the data. In general, there are two types of errors: sampling error and non-sampling error.

As mentioned in Chapter 1, sampling error is the difference between an estimate based on a sample and the corresponding value of the entire population. Sampling error in cellular probe data may be caused by bias of demographic characteristics of phone users, or may be caused by insufficient sample size.

In addition to sampling error, other types of errors may be introduced during any of the various complex operations used to collect and process collected data. For example, operations such as editing, reviewing, or keying data may introduce error into the estimates. These and other sources of error contribute to the non-sampling error component of the total error of survey estimates. Non-sampling errors may introduce random variability of the data. Therefore, it is useful to follow up the tests result with an independent data source.

2.6.2 Parametric Tests

To compare traffic information such as travel speeds or travel times from different data sources, e.g. cellular probe data, probe vehicle data and sensor data that may have different distributions, many standard tests are available. They can be categorized into parametric and non-parametric depending on the data characteristics.

In statistics literature, randomness property is essential for the theoretical base of many classical parametric statistical tests. A random process refers to a process that can produce independent and identically distributed (i.i.d) samples. If an observed value in the sequence is influenced by its position in the sequence, or by the observations which precede it, the process is not truly random. Though parametric tests are stronger than nonparametric tests in general, it is crucial to test the normality of data before applying any parametric tests. Statistics software SPSS 14.0 is used in statistical analyses.

Among the parametric tests, the basic test is to review the correlation/similarity of cellular probe data and any other data set. Correlation test describe the strength of association between variables. Chi-squared test can be used to test whether there is any association between two categories, e.g. if cellular probe data is more likely to have higher travel speed than vehicle probe data. The Fisher's exact test is used when we want to conduct a chi-squared test but one or more of the data has an expected frequency of five or less. Paired t-test can establish where one set has a significantly different mean from the other set as well. In Paired t-test, the data are naturally paired, e.g. the spot speeds measured by sensor and cellular probe technology. ANOVA, the analysis of variance is equivalent to a two sample t test, only that t is used if there are more than two samples. It is used to establish whether three or more variables have the same mean. “Statistically significant” means the difference between values of two variables is not likely to be due to random chance alone.

No conclusion on whether the two distributions are the same can be drawn based on the standard test results. However, a high probability value gives an indication of the similarity between the two sample data sets. On the other hand, a low probability value shows that the distributions are different with its given level of significance.

2.6.3 Nonparametric Tests

Traffic information is influenced by many factors, such as the time of day, day of week, road type, weather, and so on. Their randomness is questionable when we are exploring the traffic information during time interval of interest. Investigations of randomness of a given sequence often require statistical tools for comparing the distributions. Among them, goodness-of-fit tests and entropy estimates are two well-understood concepts [11], [12]. To test the normality of variables in which we are interested, One Sample Kilmogorov–Smirnov test is most widely applied. By comparing the distributions of the variables in which we are interested with a given distribution (e.g. normal distribution), p-value can be reported to indicate if there is significant difference between these two distributions.

In SPSS 14.0, a collection of nonparametric tests are available. These tests make minimal assumptions about the underlying distribution of the data. The nonparametric tests can be grouped into three broad categories based on how the data are organized:

- One sample tests that analyze one variable, e.g. binomial test, chi-square test, Kolmogorov–Smirnov test, runs test;
- Related sample tests that compare two or more variables for the same set of cases. In this study, this group of tests is the most useful one in comparing the traffic information from different data sources and concluding whether they are similar or not. Two related samples McNemar test, Sign-rank test, and Wilcoxon test are available in SPSS 14.0. Since our data are not binary,

McNemar test is not applicable here. Singed-rank test computes the differences between two variables for all cases and classifies the difference as positive, negative, or tied. If the two variables are similarly distributed, the number of positive and negative differences will differ significantly. Wilcoxon test is used to decide whether the distributions of two paired variables in two related samples are the same. It is the nonparametric version of paired-t test. The test takes into account the magnitude of the differences between two paired variables; therefore, it is more powerful than the sign test. For more than 2 related samples, Kendall test can be applied to test whether k related samples are from the same population.

- Independent sample tests analyze one variable that is grouped by categories of another variable. The Mann-Whitney test is the non-parametric version of the independent samples t-test and can be used when one does not assume that the dependent variable is a normally distributed interval variable. The Kruskal Wallis test is used when there are one independent variable with two or more levels and an ordinal dependent variable. In other words, it is the nonparametric version of ANOVA and a generalized form of the Mann-Whitney test method since it permits two or more groups.

2.7 Summary

In this chapter, we reviewed data quality measures, existing similar projects and evaluation data sources applied in similar evaluations. An evaluation procedure was developed and the sampling method was selected based on the identified important factors that may influence the cellular probe data quality. Various statistical tests were reviewed. It was addressed that the usage of these tests depends on the data properties. Normality tests are necessary before selecting the parametric or nonparametric tests.

3 Data Summary

Four groups of data are used in this study. These include cellular probe data, vehicular probe data, sensor data (CHART data) and Baltimore road network data. The provider and data format are shown in Table 3-1. The details of the data are discussed in the following sections.

Table 3-1
Data sources

Agency/Provider	Content	File type
MMTIS	Cellular probe data	Text files
	Road network	Shape files
MotionMaps	Vehicular Probe Data (Raw data)	Text files
	Segmented data for each run	Database files
	Description data	PowerPoint files
CHART	Spot traffic speed data	Excel file
	Spot traffic flow data	Excel file

3.1 Road Network Data

The Baltimore road network is provided by ITIS Holdings in RDS-TMC (Radio Data system- Traffic Message Channel) network system format. The database files of nodes and links are used in this evaluation project. Each node in the TMC network file is associated with road name, positive offset node, negative offset node and coordinates. This information is essential to perform further study using the cellular probe data and the vehicular probe data. Based on the sequence of nodes visited in a probe vehicle survey run, the direction of the link can be found to match the corresponding traffic flow characteristics in the cellular probe data. Table 3-2 shows an example of the TMC node file.

The entire road network contains 637 nodes and 1274 directed links. Figure 3-1 illustrates the Baltimore road network.

Table 3-2
An example of the TMC road network file

ID	NUM	FIRST_ NAME	OFFSET	OFFSET0	LATTITUDE	LONGITUDE	X	Y
04365	I-97	MD-178/ EXIT 5	4364	4366	003904264	-07661521	-76.62	39.04
04366	I-97	MD-32/ EXIT 7	4365	4367	003905869	-07663390	-76.63	39.06
04367	I-97	MD-3/ EXIT 7	4366	4368	003906920	-07663844	-76.64	39.07

3.2 Cellular Probe Data

3.2.1 Data Format

The cellular data is provided in 5-minute intervals. An example of cellular probe data is shown in Table 3-3. In each record, there are 6 variables:

- timestamp: Each time stamp represents a 5-minutes interval. For example, 1/26/2006 12:05:00 AM means the traffic flow during 1/26/2006 12:05:00 AM to 1/26/2006 12:09:59 AM will have the same speed and travel time on the current link.
- TMCLinkID: TMC link identification number. The link identification number is represented by the destination node number of a link. For example, TMCLinkID 4368 in direction 1 represents the link from node 4367 to 4368.
- Direction: the direction of the link, the direction of a link is decided based on the TMC road network file. When the previous node is defined as positive offset node in the TMC road net work file, the link is in direction 1. When the previous node is defined as negative offset node, the link is in direction 0.
- Speed: the average speed of traffic flow on current link in kilometer per hour;
- journeytime: the average travel time of traffic flow on the current link in seconds;
- Length: the length of the current link in meter.

The units used in cellular probe data are in metric system. The metric units are transformed into English units in data comparison.

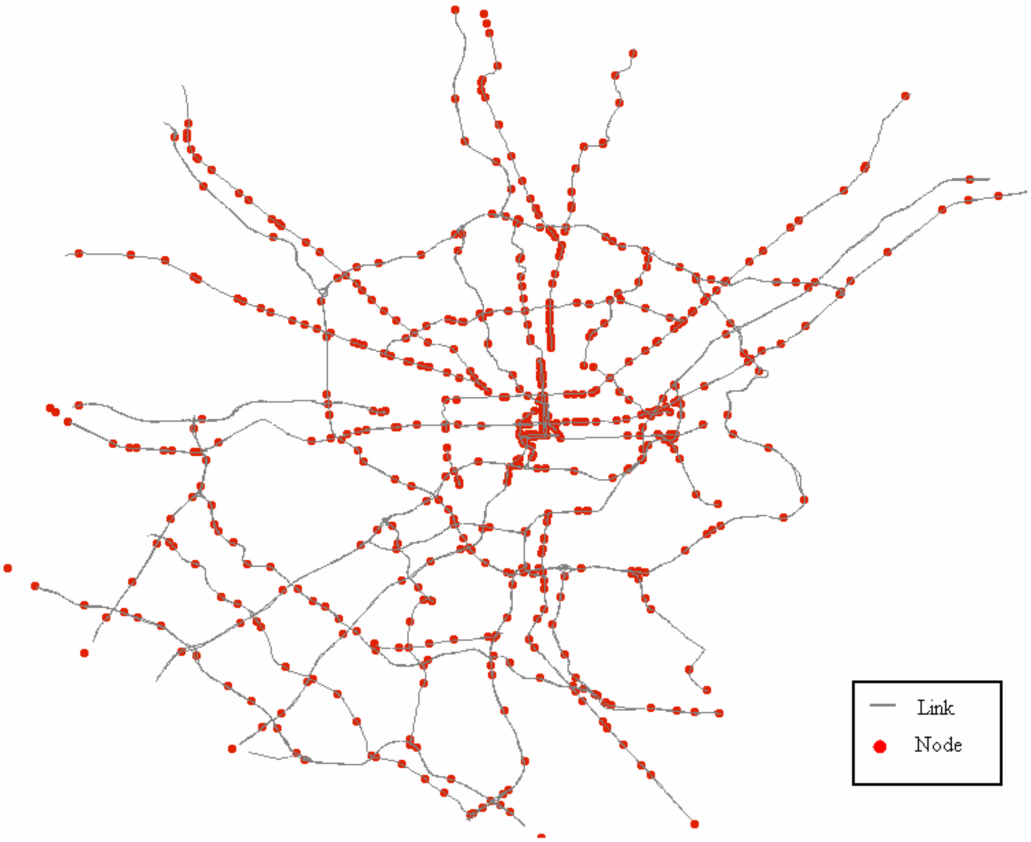

Figure 3-1
Baltimore TMC road network cellular data

Table 3-3
An example of cellular probe data

timestamp	TMCLinkID	Direction	Speed(kph)	journeytime(secs)	Length(m)
1/26/2006 12:05:00 AM	4368	1	80	129.96	2888
1/26/2006 12:10:00 AM	4368	1	80	129.96	2888
1/26/2006 12:15:00 AM	4368	1	80	129.96	2888

3.2.2 Data Coverage

The cellular probe data is provided within a 21 days interval, ranging from Jan 26th, 2006, 12:00 AM to Feb 16th, 2006, 11:55 PM. The cellular probe data covers 596 nodes and 1107 links in total. There are 553 links in direction 0 and 554 links in direction 1. The spatial coverage of the cellular probe data is shown in Figure 3-2. It should be noted that the uncovered data are mainly those node on the boundary of the network. Since those nodes are not included in the vehicular probe survey route, they will not affect the comparison between vehicular probe data and cellular probe data.

3.3 Vehicular Data

3.3.1 Data Format

The vehicular probe data is provided by MotionMaps based on survey routes. Each route is associated with three types of files:

- Raw GPS data file: which records the coordinates of probe vehicle in every second;
- Excel file: which represents the routes in the Baltimore road network and the summary files for each node and each run. Table 3-4 shows an example of the raw data file.
- Summary files: in which the trajectories of the vehicles were derived from the raw data and matched to the locations in the road network. Table 3-5 illustrates an example of summary file based on each run of survey. Each record is composed of 22 fields that include the run number, the name of the location, the start time of each run, the start time from each node, the travel time on the current link, the average travel speed on the current link, etc. In these summary files, the coordinates in raw GPS data are matched to the TMC node numbers or detailed road locations. The data analysis is based on these Excel summary files rather than the text files of raw data.

3.3.2 Data Coverage

The probe vehicle survey was conducted by MotionMaps (search and confirm the name) on 26, 27 and 30 January and 2 and 3 February 2006. Table 3-6 lists the detailed survey routes and survey times. Though the survey times spanned from early morning (6 AM) to evening (9 PM), most of the surveys were focused on the morning rush hours and afternoon rush hours.

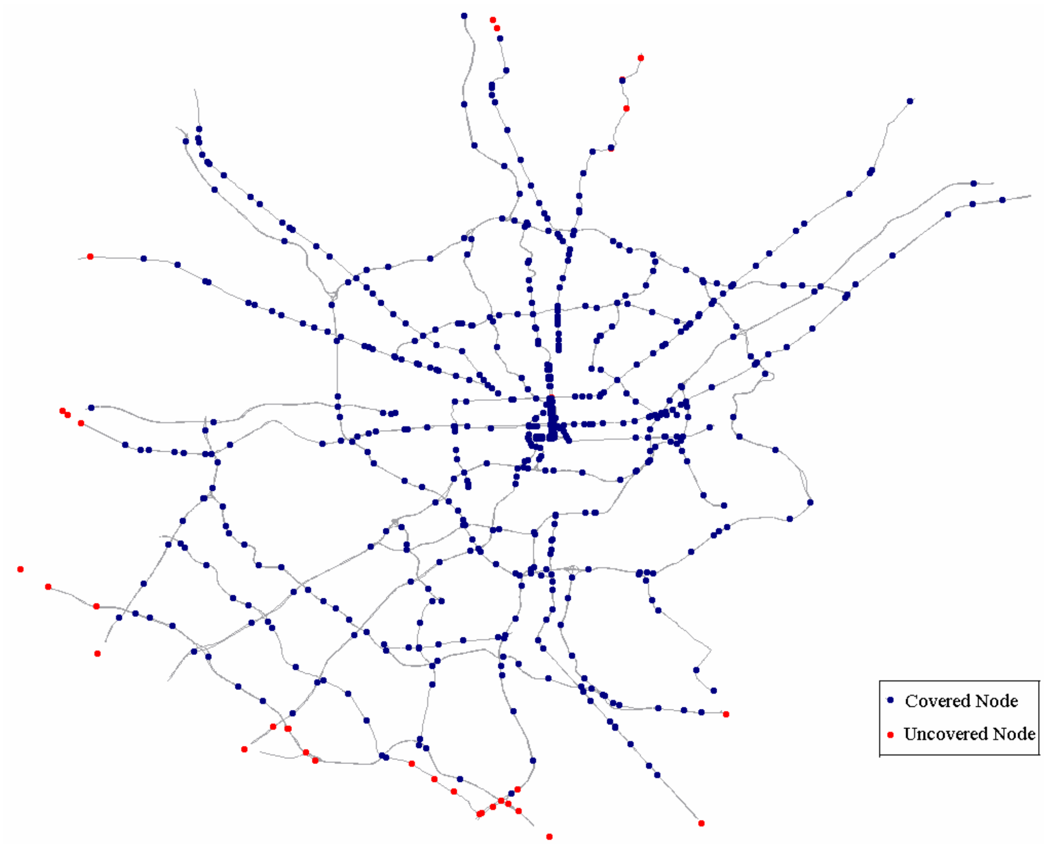

Figure 3-2
Spatial coverage of cellular probe data

Table 3-4
An example of the probe vehicles' raw GPS data

$G,012606,152353,+3908.469,-07650.639,000.00
$G,012606,152354,+3908.469,-07650.639,000.00
$G,012606,152355,+3908.469,-07650.639,000.00
$XA,012606,152355
$XD,012606,152355
$G,012606,152359,+3908.469,-07650.639,000.00

Table 3-5
An example of the probe vehicles summary file

PXTOT_ID	XL_ID	XL_NAME	XL_TYPE	XL_POSTEDS	START_TIME	DATE	…
5	1	Marriotsville Rd Ram	Start	65	155800	20060126	…
13	1	Marriotsville Rd Ram	Start	65	161904	20060126	…
21	1	Marriotsville Rd Ram	Start	65	164114	20060126	…
29	1	Marriotsville Rd Ram	Start	65	170409	20060126	…

The survey routes cover the major freeways and arterials in Baltimore network that include road segments on route I-70, I-95, I-395, I-695, I-895, MD-40, US-1, MD-45, MD-140,MD-146, and Martin Luther King Boulevard. There are 10 to30 runs for each route with a total of 500 runs of probe vehicles. The length of each survey run ranges from 1 mile to 30 miles.

The survey time of these runs covered weekend and weekdays, morning peak hours, afternoon peak hours and non-peak hours.

Figure 3-3 represents these survey routes. Table 3-6 lists all probe vehicle survey routes and survey dates.

3.4 Sensor Data

3.4.1 Data Format

The CHART (Coordinated Highways Action Response Team) is a multi-jurisdictional and multi-disciplinary program started in mid-1980's in Maryland. The mission of CHART includes incident management, traffic and roadway monitoring, traveler information and traffic management. Its coverage area is not limited to the Baltimore-Washington corridor, but expanded into a statewide program. Nowadays, traffic speed detectors are deployed along 155 centerline miles of the heaviest traveled freeways. These detectors provide average speed and density information along a segment of roadway. This program is enhanced by a newly constructed command and control center called the Statewide Operations Center (SOC). SOC is functioning 24 hours-a-day, seven days a week with satellite Traffic Operations Centers (TOCs). CHART traffic data is used in this evaluation study.

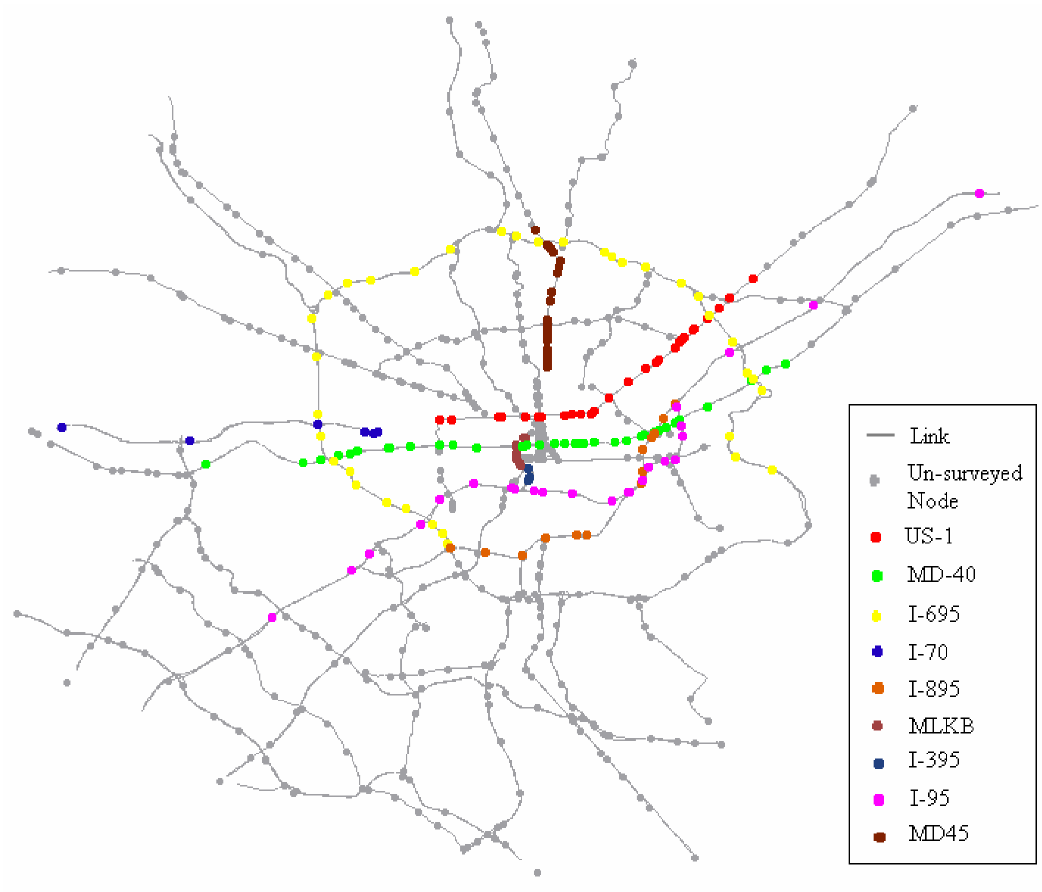

Figure 3-3
Survey routes of probe vehicles

Table 3-6
Details of probe vehicle survey routes

Route Number	Route Name	From	To	Thursday Jan 26th	Friday Jan 27th	Monday Jan 30th	Thursday February 2nd	Friday February 3rd
I-070E	Baltimore National Pike	Marriotsville Road	Security Boulevard	X				
I-070W	Baltimore National Pike	Security Boulevard	Marriotsville Road	X				
I-095N	I-95	I-195, Exit 47	MD-43 White Marsh Blvd, Exit 67			X		X
I-095S	I-95	MD-43 White Marsh Blvd, Exit 67	I-195, Exit 47			X		X
I-395N	I-395 and ramp to/from North	I-95/Exit 55	Conway St and/or ML King Blvd					X
I-395S	I-395 and ramp to/from North	Conway St and/or ML King Blvd	I-95/Exit 55					X
I-395N	I-395 and ramp to/from North	I-95 Exit 55	Martin L. King, Jr. Blvd to Eutaw					X
I-395S	I-395 and ramp to/from North	Martin L. King, Jr. Blvd to Eutaw	I-95 Exit 55					X

I-395N	I-395 and ramp to/from South	I-95 Exit 50	W. Conway St	X				X
I-395S	I-395 and ramp to/from South	W. Conway St	I-95 Exit 50	X				X
I-395N	I-395 and ramp to/from South	I-95 Exit 50	Martin L. King, Jr. Blvd to Eutaw					X
I-395S	I-395 and ramp to/from South	Martin L. King, Jr. Blvd to Eutaw	I-95 Exit 50					X
I-695IL	Baltimore Beltway	I-95/Exit 11	Reistertown Road, Exit 20		X			X
I-695OL	Baltimore Beltway	Reistertown Road, Exit 20	I-95/Exit 11		X			X
I-695IL	Baltimore Beltway	Reistertown Road, Exit 20	US 40 Pulaski Hwy Exit 38		X			
I-695OL	Baltimore Beltway	US 40 Pulaski Hwy Exit 38	Reistertown Road, Exit 20		X			
I-695IL	Baltimore Beltway	Reistertown Road, Exit 20	Cove Ave Exit 41					X
I-695OL	Baltimore Beltway	Cove Ave Exit 41	Reistertown Road, Exit 20					X
I-895N	Harbor Tunnel Thruway (+I-95)	I-695 Baltimore Beltway	I-95 MD-43 White Marsh Blvd, Exit 67			X		
I-895S	Harbor Tunnel Thruway (+I-95)	I-95 MD-43 White Marsh Blvd, Exit 67	I-695 Baltimore Beltway			X		
MD-026N	Liberty Heights Ave	Fulton Ave	Washington Ave				X	

MD-026S	Liberty Heights Ave	Washington Ave	Fulton Ave			X	
MD-045N	York Road	East 43rd Street	MD-131/Seminary Ave			X	
MD-045S	York Road	MD-131/Seminary Ave	East 43rd Street			X	
MD-140N	Reistertown Rd	Fulton Ave	Greenspring Valley Road			X	
MD-140S	Reistertown Rd	Greenspring Valley Road	Fulton Ave			X	
MD-146N	Dulaney Valley Road	MD-45 York Road/Joppa Road	East Seminary Ave			X	
MD-146S	Dulaney Valley Road	East Seminary Ave	MD-45 York Road/Joppa Road			X	
US-001N	North Ave/Belair Rd	Hilton Parkway	Silver Spring Road			X	
US-001S	Belair Rd/North Ave	Silver Spring Road	Hilton Parkway			X	
US-040E	Nat. Pike/Edmondson/Mulberry	Rogers Ave	MD 139 Charles St		X		
US-040W	Franklin/Edmonson/Nat. Pike	MD 139 Charles St	Rogers Ave		X		
US-040E	Orleans St/Pulaski Hwy	MD-2/St Paul St	Rossville Boulevard		X		
US-040W	Pulaski Hwy/Orleans St	Rossville Boulevard	MD-2/St Paul St		X		
	Martin Luther King, Jr. Blvd	Washington Blvd	Eutaw St	x			X

	Martin Luther King, Jr. Blvd	Eutaw St	Washington Blvd	x	X
	Pratt St (eastbound)	M.L. King, Jr. Blvd	President St		X
	Lombard St (westbound)	President St	M.L. King, Jr. Blvd		X
	Pratt St (eastbound)	M.L. King, Jr. Blvd	President St		X
	Lombard St (westbound)	President St	M.L. King, Jr. Blvd		X

RTMSs (Remote Traffic Microwave Sensor) are general-purpose detectors, which can operate in all weather conditions, and can measure presence, volume, occupancy, speed and classification information. Typically, these sensors are mounted facing perpendicular to a roadway. RTMS devices are capable of measuring the road speed in several lanes of a multi-lane roadway, and typically the reports from these sensors are averaged over these lanes. The range and speed of these objects can then be calculated by measuring the time lag of reflected microwave signals.

CHART traffic data is downloaded as a Microsoft Excel file from the website of the Center for Advanced Transportation Technology Laboratory (CATT LAB) in University of Maryland, College Park. Table 3-7 shows an example of downloaded data format. In general, the average spot speeds of traffic are available in 5-minute intervals. In some cases, this interval can be 10 minutes or even longer. Different from cellular probe data, time intervals may not start exactly from 0:00 everyday. When comparing the travel speed, the sensor data should be matched to the cellular data whose time interval contains the sensor timestamp.

Table 3-7
An example of sensor data

UMD CATT Lab Traffic Data				
TIME	DIRECTION	LOCATION	SPEED	VEHICLES/HOUR
1/26/2006 0:21	Inner Loop	I-695 @ I-70 Inner Loop	57	1224
1/26/2006 0:26	Inner Loop	I-695 @ I-70 Inner Loop	59	1284
1/26/2006 0:31	Inner Loop	I-695 @ I-70 Inner Loop	60	1620
1/26/2006 0:36	Inner Loop	I-695 @ I-70 Inner Loop	58	1452
1/26/2006 0:41	Inner Loop	I-695 @ I-70 Inner Loop	55	1356

3.4.2 Data Coverage

On Interstate 695 there are 15 locations. The locations of these sensors are shown in Figure 3-4. Based on the availability of sensor data, we selected 10 locations for data quality reference. The two between Stevenson Road and Greenspring Avenue cannot be matched to any TMC node numbers. The two at the junction of Interstate 695 and Interstate 95 did not have traffic data available during the time interval of interest (26 January to 3 February 2006). The location of the sensor at junction I-695 and Providence was not matched to any TMC node either.

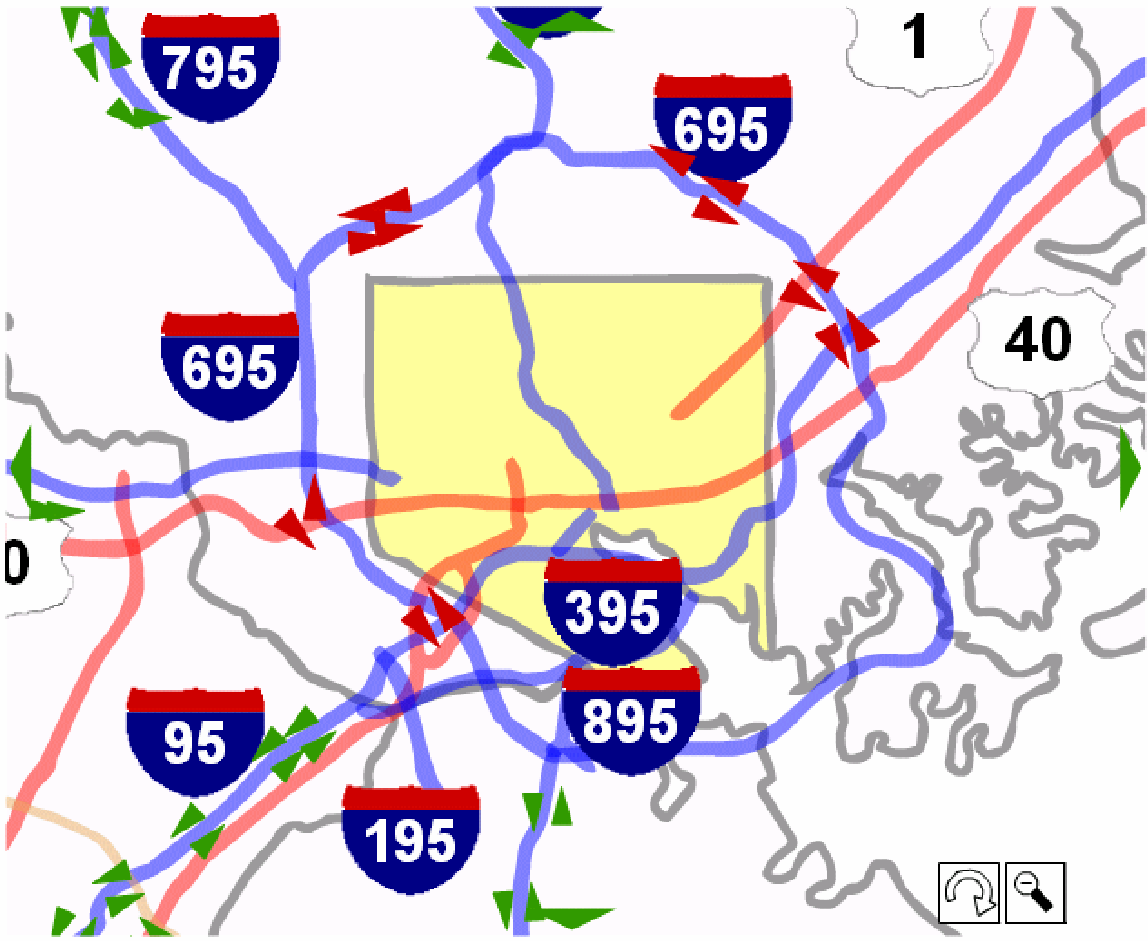

Figure 3-4
Sensor locations on I-695

The probe vehicle survey started from 26 January 2006 and finished on 3 February 2006. Therefore, the CHART traffic speed data on corresponding locations are collected for evaluation purpose. Table 3-8 shows the summary of the time duration on each sensor location. At locations I-695 @ US 40W Outer Loop and I-695 @ I-70 Inner Loop, the data is not available from 31 January to 3 February 2006. At junction I-695 @ Stevenson Road Outer Loop, the sensor data is not available from 26 January to noon time of 27 January 2006.

Table 3-8
Summary of sensor data time interval

Location	StartTime	EndTime
I-695 W (O/L) US 40 W Outer Loop	1/26/2006 0:17	1/31/2006 0:09
I-695 @ I-70 Inner Loop	1/26/2006 0:21	1/31/2006 0:07
I-695 EB @ Stevenson Rd Inner Loop	1/26/2006 0:35	2/3/2006 23:45
I-695 WB @ Stevenson Rd Outer Loop	1/27/200611:47	2/3/2006 23:50
I-695 EB @ Joppa Rd, Ex 29 Inner Loop	1/26/2006 0:55	2/3/2006 23:55
I-695 WB @ Joppa Rd, Ex 29 Outer Loop	1/26/2006 0:55	2/3/2006 23:55
I-695 WB @ US 1, Ex 32 Outer Loop	1/26/2006 0:50	2/3/2006 23:50
I-695 EB @ US 1, Ex 32 Inner Loop	1/26/2006 0:55	2/3/2006 23:55
I-695 WB at I-95, Ex 33 Outer Loop	1/26/2006 0:52	2/3/2006 23:57
I-695 EB at I-95, Ex 33 Inner Loop	1/26/2006 0:52	2/3/2006 23:57

Traffic information derived from sensor data cannot be deemed as the “true” traffic condition on road, due to its accuracy limitations. It is used to serve the data comparison purpose in this project in providing a direct reference for cellular probe data quality.

3.5 Summary

In this chapter we summarized the format, spatial and temporal coverage of cellular probe data, vehicular probe data and sensor data. This summary helps us understand the relation between these data sources and addresses the potential problems in data processing and data quality evaluation.

4 Link Traffic Information Evaluation

4.1 Link Length Comparison

When comparing the link lengths in cellular data and vehicular data, it was observed that for the same link, these two lengths were sometimes different. This deviation is caused by the labeling system in the TMC road network. An example is shown in Figure 4-1. TMC Link 4481 in direction 0 represents the road segment from node 4482 to node 4481. In cellular probe data file, the length of this link is 1460 meter or 0.9 mile. The same segment in probe vehicle data file is 1.53 mile. The discrepancy comes from the location of node 4482. In cellular data, a link typically extends from the last on/off ramp at the preceding junction (TMC location) to the last on/off ramp at the next junction (TMC location) on an Interstate roadway. Therefore, measuring a road segment from one TMC point location to the next will give a value that is often different from the actual distance. Though the location of the TMC nodes is a guide to the location and is usually positioned close to the center of the junction, on Interstate roads that have long on/off ramps, the actual location of the TMC node can be significantly away from the center of the junction.

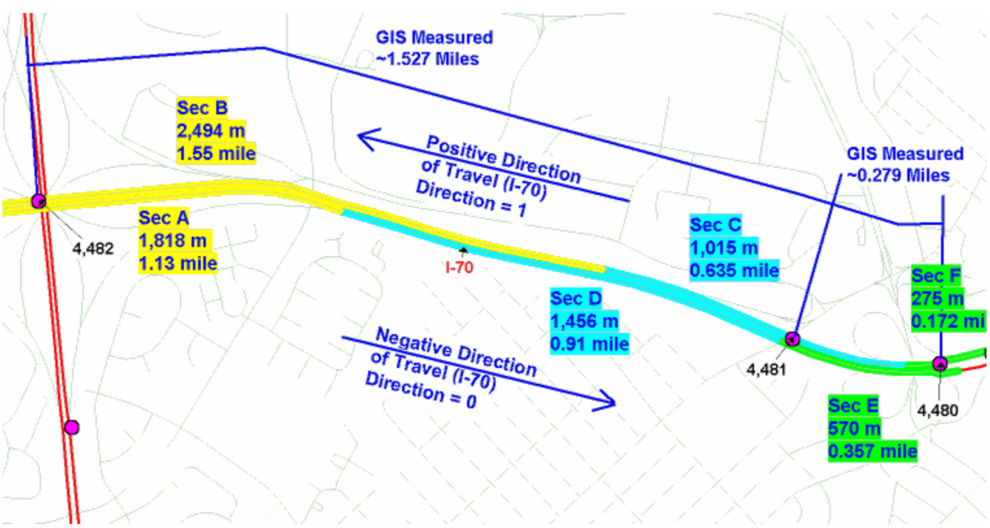

Figure 4-1
An example of link length discrepancy

Table 4-1 summarizes the deviation of the link lengths in cellular data from the link lengths in vehicular data. The average link lengths in these two data sets are almost the same. The discrepancy is 0.35 percent of the average vehicular link length and the average absolute error is 19.8 percent of the average vehicular link length. The histogram analysis shows that 31.77 percent of links have discrepancies between 20 to 50 percent of the corresponding vehicular link length, and 5.39 percent of links have a discrepancy more than 50 percent.

Table 4-1
Summary of link length discrepancy

% of Abs. Error			Frequency		Probability Density	
[0, 10%)			2568		49.24%	
[10%, 20%)			709		13.60%	
[20%, 50%)			1657		31.77%	
[50%, 100%)			268		5.14%	
>=100%			13		0.25%	
	Cell. Length (mile)	Veh. Length (mile)	Error (mile)	% Error	Abs Error (mile)	Abs. % of Error
Avg.	0.846	0.849	-0.003	0.35	0.168	19.8

Any link length discrepancy may cause a corresponding difference in link travel time, especially when the length of road is short. Because of these discrepancies it is inappropriate to compare link travel time. Instead, the travel speed is a better indicator of traffic condition on links. When the road segment is long, the on/off ramps usually compensate each other and the total lengths of links in cellular data and vehicles data have smaller discrepancies. Traffic speeds on a path usually vary from segment to segment; when a path is long enough and the path discrepancy is relatively small, travel time is a better indicator of traffic condition.

4.2 Data Processing

To compare travel speed on links EXCEL Marcos were developed to prepare the data. The following major steps are taken in the data preparation procedure:

Step 0: Combine and sort segmented probe vehicle data for each survey run;
Step 1: Catch exceptions such as nodes without matching node number or potential error in matched node number, in the probe vehicle data;

Step 2: Query road network data to define the direction for each link;
Step 3: Transfer date and time in the probe vehicle data to the same format as in the cellular probe data;
Step 4: Develop an ACCESS project to find the neighboring data records in the cellular probe data for each probe vehicle data record.
Step 5: Calculate the interpolated travel speeds from query results.

In this procedure, the date and time information are transferred to numbers to simplify calculations and queries, e.g. 6 AM on 27 January 2006 will be transferred to 38744.25, and each 5 minutes equals to 5/24/60=0.003472. Therefore the query for neighboring cellular data will based on the TMC link ID, link direction and the value of data and time. For each probe vehicle record with valid TMC link ID and link direction, two cellular records are found: the interval includes the start time and the interval next to it.

The reason to obtain cellular data in two intervals is the need for interpolation. When the probe vehicle starts from one 5-minute interval but ends in another 5-minute interval, interpolation is needed. In all cases where the cellular and vehicular link speed pairs can be matched, there are only 4 pairs have a travel time longer than 10 minutes and need to be treated separately. All the other travel times on a link are within a 10-minute interval and the data needed for interpolation can be found in the two neighboring cellular records.

The interpolation depends on the time the vehicle spends in each interval. For example, a probe vehicle starts from node 4520 at 7:19:32 on 27 January 2006, and arrives at node 4521 at 7:20:22. The duration is 50 seconds and the distance traveled by the vehicle is 0.86 mile. The probe vehicle speed on this link is:

$$Speed_{veh} = 3600 * length_{veh} / (t_{end} - t_{start}) = 3600 * 0.86 / 50 = 61.92 \text{ mile per hour} \quad (4\text{-}1)$$

To find the matching cellular travel speed on the same link, two records will be retrieved by the query (as shown in Table 4-2).

Table 4-2
An example of travel speed interpolation: query result

timestamp	TMCLinkID	Direction	Speed (kph)	Journeytime(secs)	Length(m)
7:15:00	4521	1	98.07	34.69	945
7:20:00	4521	1	93.8	36.27	945

The interpolated cellular speed is calculated as follow:

$$\begin{aligned} Speed_{cell} &= ((stamp_2 - t_{start}) * Speed_1 + (t_{start} - stamp_2) * Speed_2) / ((t_{end} - t_{start}) * 1.609) \\ &= 98.07 + 22 * 93.8) / (50 * 1.609) = 59.78 \text{ mile per hour} \end{aligned} \quad (4\text{-}2)$$

In this example, the interpolated cellular link speed is 59.78 mph and the vehicular link speed is 61.92 mph. The cellular link speed is 2.14 mph less than the vehicular one. This difference is 3.5 percent of vehicular link speed. When comparing travel times, the cellular travel time is about 35.38 seconds and the vehicular travel time is 50 seconds. The difference is 14.22 seconds and this discrepancy is more than 28 percent of the vehicular travel time. The difference of these two results is caused by the link length discrepancy. The link length in cellular data is 945 meter or 0.59 mile. This value is much shorter than the distance traveled by the probe vehicle (0.86 mile). The example illustrates that it is more reasonable to compare travel speeds on links rather than travel times.

4.3 Data Quality Measures

Link speed data quality measures are based on the data obtained by the probe vehicles. The actual travel speed calculated from the GPS recordings is deemed as “Ground Truth”, and is compared with the corresponding speed reported from the ITIS cellular traffic system (CFVD™ Speed).

Statistical sampling methods suggest that roughly 4% to 5% of probe vehicle penetration can provide a good estimate of travel times based on simulation results [1]. In our study, the number of probe vehicles is far less than the recommended values due to budget limitations. Therefore the vehicle probe data that are used in this study may not accurately represent the ““Ground Truth”.” However, the values of data quality measures reported are still based on the “Ground Truth” assumption.

When comparing these two travel speeds, the system error is calculated as following:

Error = (CFVD™ Speed- “Ground Truth”) (4-3)
Percentage of Error = 100*(CFVD™ Speed- “Ground Truth”)/ “Ground Truth” (4-4)

The following measures are selected: 1) Average Error, 2) Average Percentage of Error, 3) Average Absolute Error, and 4) Percentage of Average Absolute Error. These measures are calculated for the entire data set and for specific routes, time of day, and level of service. The statistical tests results are reported in each comparison scenario.

4.4 Summary of Statistical Tests

4.4.1 Evaluations by Route

9 routes are analyzed in this section. Table 4-3 summarizes the speed deviations for these. The average error ranges from 0.19 mile per hour to 23.59 mile per hour. It is observed

that the average errors on freeways (I-895, I-695, I-395, I-95 and I-70) are much smaller than those on arterials (US-1, Martin Luther King Boulevard, MD 40 and MD 45). The 50 percentile of error (50%kv, median) shows the same trend. The 85 percentile of error (85%kv) represents the range of error for most records. The average absolute errors show the same trend. It is observed that the average percentages of absolute error on arterials are extremely high. One contributing reason for high percentages is that some probe vehicular link speeds are very low (5-10 miles per hour). When there are 10 to 20 miles speed deviations, the percentage of absolute errors are around 100% to 200%.

Table 4-3
Summary of speed deviations

	Avg. Cell Speed mi/h	Avg. Veh. Speed mi/h	Avg. Error mi/h	Avg. Abs. Error mi/h	50%kv Abs. Error mi/h	85%kv Abs. Error mi/h	Avg. % of Error	Avg. % of Abs. Error
I-70	49.63	53.27	-3.63	8.04	7.56	12.42	-3.99	15.27
I-95	57.50	55.85	1.65	9.82	6.20	16.67	50.63	63.54
I-395	45.52	48.03	-2.51	12.36	9.72	22.12	-3.77	25.55
I-695	55.24	55.43	-0.19	8.10	5.77	14.97	7.28	21.17
I-895	55.47	52.37	-3.09	10.30	6.36	16.67	29.85	42.04
US-1	47.73	24.14	23.59	24.37	14.83	31.71	148.84	151.21
MD-40	42.32	27.94	14.37	17.39	24.94	33.75	87.12	96.70
MD-45	43.57	20.43	23.14	23.84	21.71	38.66	164.87	167.02
MLKB	38.83	21.45	17.38	22.64	24.14	35.73	147.29	163.67

Figure 4-2 illustrates the comparison of average cellular speeds and average vehicular speeds on each route. It is clear that on all arterials, the average cellular speed overestimates the average vehicular speed. On freeways, the differences between the average cellular speed and the average vehicular speed are less than 5 miles per hour. The average cellular speeds underestimate vehicular average on I-70, I-395 and I-695, while overestimate on I-95 and I-895. Figure 4-3 illustrates the comparison of the percentage of average error and the percentage of average absolute error on each route. On I-70, I-395 and I-695, the percentage of average error is much smaller than the percentage of average absolute error. This indicates that the deviation between cellular speeds and vehicular speeds are relatively evenly spread across the axis. On US Route 1, MD-40, MD-45 and Martin Luther King Boulevard, average percentage of errors and average percentage of absolute errors are consistently large which indicates that cellular speeds consistently overestimate vehicular speeds. As shown in Table 4-3, the average absolute errors on freeways are around 10 mile per hour and those on arterials are around 20 miles per hour, while the range of the average percentage of absolute errors on arterials are much larger. This is because that the vehicular speeds are the denominators in the calculation of percentages. The low vehicular link speeds on arterials bring high percentage of absolute error.

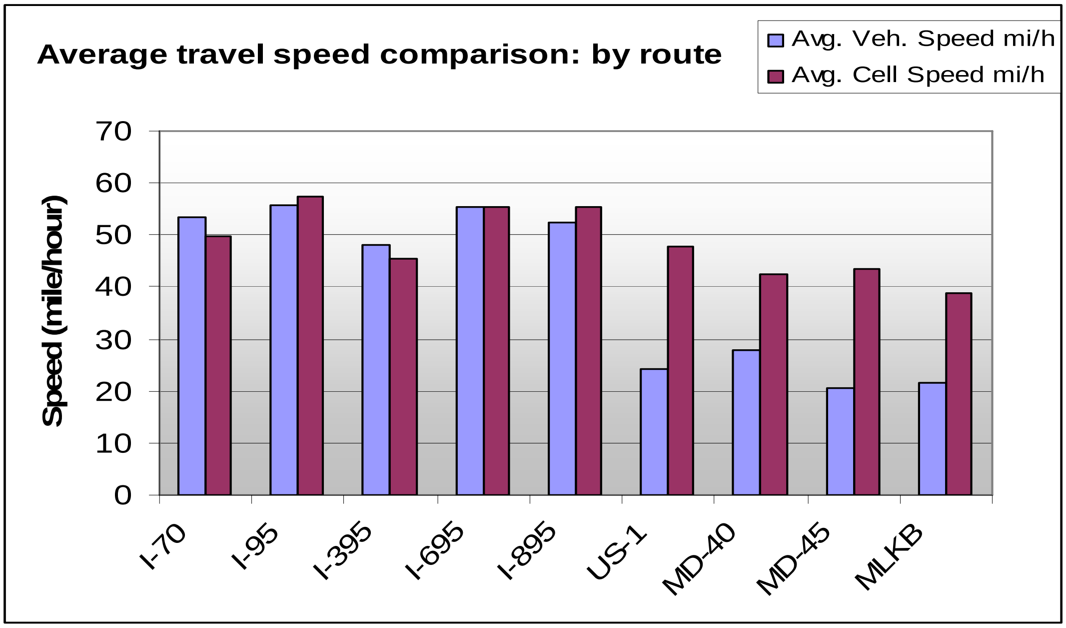

Figure 4-2
Comparison of average speeds: by route

These extremely high percentages of absolute errors can dramatically degrade the data quality measures. As an example, if 95% of the data records have an average absolute error less than 10% of the “Ground Truth,” and 5% of them have an average absolute error above 1000% of the “Ground Truth,” the quality of estimation is impressively good. However, when calculating the average absolute error of all records, the average absolute error of all records might be more than 60% of the “Ground Truth,” which makes the estimation method look not promising at all. Therefore, it is important to do histogram analyses for these measures as well. It is important to refer to the corresponding histogram when reading the values of data quality measures in order to better understand and evaluate the data quality. In the following link traffic information and path traffic information evaluation, the quality measures will be presented associated with histogram analysis.

To explore the average percentage of absolute error, we categorize the percentage of absolute error in 5 ranges: [0, 10%), [10%, 20%), [20%, 50%), [50%, 100%), and [100%, ∞). The percentage of records that fall in each range on each route is shown in Figure 4-4. Comparing Interstates with arterials, the percentages of records in the first two categories are significantly larger than those of latter.

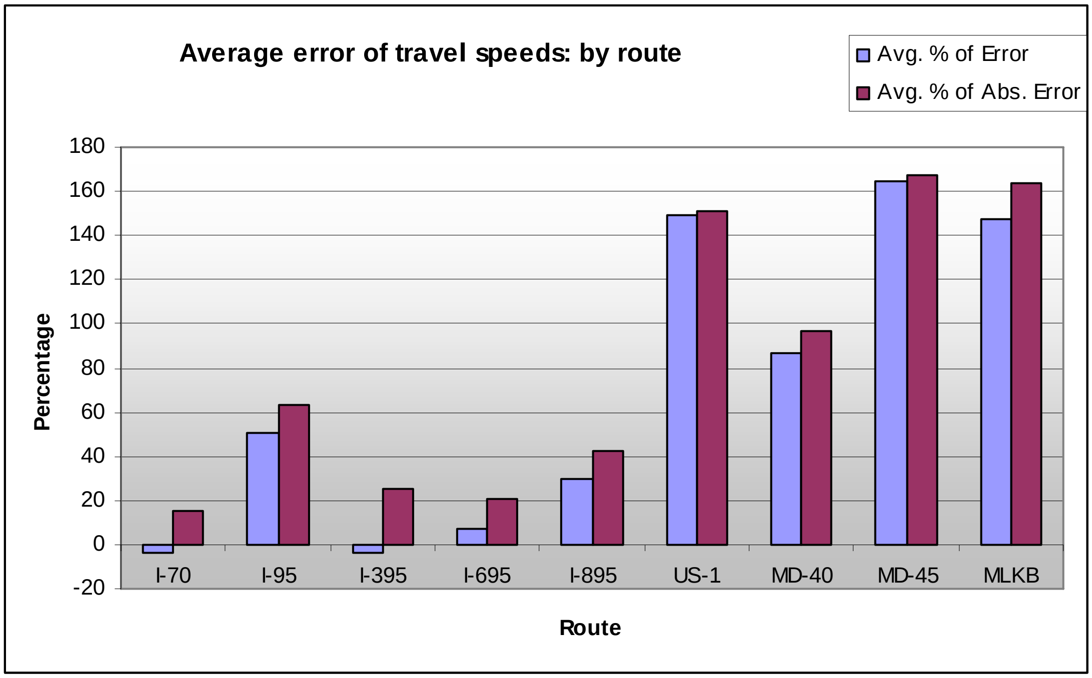

Figure 4-3
Comparison of average error: by route

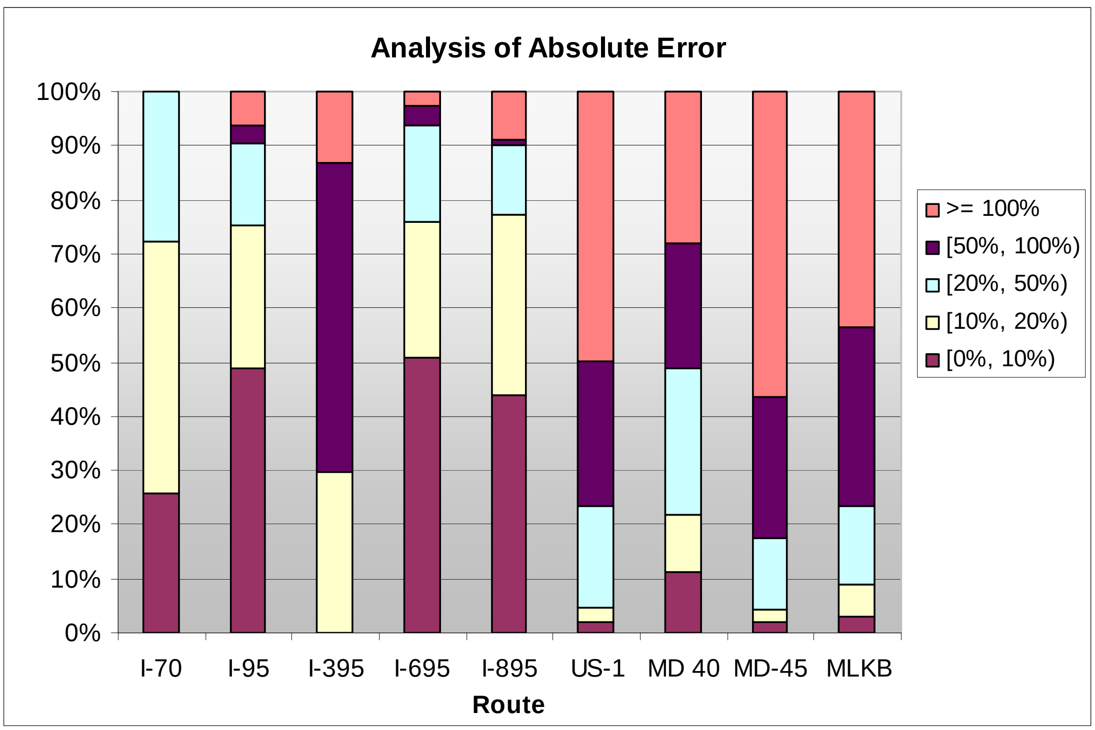

Figure 4-4
Analysis of percentage of absolute error

Compared to other Interstates, I-95 has the highest average percentage of absolute error, though more than 75% of the records fall in the range of [0, 20%). There are some records with extreme low speeds, e.g. 5 miles per hour that the cellular probe data does not catch. These records contribute a lot to the average percentage of error. I-395 has the highest percentage of records with high absolute error among Interstates. This is partly due to the fact that the north part of the survey route on I-395 is moving towards inner harbor and the close on and off ramps have significant impact on the main stream traffic. The detailed test results are discussed in the following sections.

4.4.1.1 I-70

58 pairs of cellular vehicle speeds are obtained on I-70 for this comparison. The summary of deviation is shown in Table 4-3. The average cellular speed underestimates the average vehicular speed by 3.63 miles per hour and the average absolute error is 15.27%. More than 72.41% of records have an absolute error less than 20 percent. 27.59% of records have an absolute error between 20 to 50 percent.

The One Sample Kolmogrov-Smirnov test results show that vehicular speeds are normally distributed but cellular speeds are not (Table 4-4). Therefore, non-parametric tests, Friedman test, Kendall's W test and Wilcoxon Signed test, are performed. The results shown in Table 4-5 indicate that the speeds from these two groups are not from the same population. In another work, statistically, these two groups of data are not similar to each other (p=0.05). The use of speed limits as a default speed value may account for this difference.

4.4.1.2 I-95

1221 pairs of cellular vehicle speeds are obtained on I-95 for this comparison. Table 4-3 shows the average cellular speed overestimates the average vehicular speed by 1.65 miles per hour and the average absolute error is 63.54%. More than 75.27% of records have an absolute error less than 20 percent of the corresponding vehicular speed. 15.07% of records have an absolute error between 20 to 50 percent.

The One Sample Kolmogrov-Smirnov test results show that neither type of speeds is normally distributed (Table 4-6, p=0.05). Non-parametric tests results shown in Table 4-7 indicate that these two groups of data are similar to each other.

Table 4-4
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		59	59
Normal Parameters(a,b)	Mean	49.6340	53.2683
	Std. Deviation	7.09855	11.56369
Most Extreme Differences	Absolute	.331	.126
	Positive	.331	.102
	Negative	-.211	-.126
Kolmogorov-Smirnov Z		2.543	.966
Asymp. Sig. (2-tailed)		.000	.309

a Test distribution is Normal.
b Calculated from data.

Table 4-5
Test statistics

N(a)	59	N(b)	59		Cellular-Vehicular
Chi-Square	8.966	Kendall's W(b)	.152	Z(c)	-2.846(a)
Df	1	Chi-Square	8.966	Asymp. Sig. (2-tailed)	.004
Asymp. Sig.	.003	Df	1		
		Asymp. Sig.	.003		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

Table 4-6
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		1221	1221
Normal Parameters(a,b)	Mean	57.5016	55.8495
	Std. Deviation	8.99168	14.70810
Most Extreme Differences	Absolute	.126	.235
	Positive	.125	.120
	Negative	-.126	-.235
Kolmogorov-Smirnov Z		4.415	8.225
Asymp. Sig. (2-tailed)		.000	.000

a Test distribution is Normal.
b Calculated from data.

Table 4-7
Test statistics

N(a)	1221	N(b)	1221		Cellular-Vehicular
Chi-Square	2.130	Kendall's W(b)	.002	Z(c)	-1.454
Df	1	Chi-Square	2.130	Asymp. Sig. (2-tailed)	.146
Asymp. Sig.	.144	Df	1		
		Asymp. Sig.	.144		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

4.4.1.3 I-395

91 pairs of cellular vehicle speeds are obtained on I-395. The average cellular speed underestimates the average vehicular speed by 2.51 miles per hour and the average absolute error is 25.55%. More than 29.67% of records have an absolute error less than 20 percent. 57.14% of records have an absolute error between 50 to 100 percent. The One Sample Kolmogrov-Smirnov test results show that vehicular link speeds are normally distributed, but cellular link speeds are not (Table 4-8). Non-parametric tests results (Table 4-9) indicate that these two groups' speeds are from the same population. That is, statistically, these two groups of data are similar to each other.

Table 4-8
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		91	91
Normal Parameters(a,b)	Mean	45.5183	48.0314
	Std. Deviation	14.68127	4.86290
Most Extreme Differences	Absolute	.237	.072
	Positive	.157	.043
	Negative	-.237	-.072
Kolmogorov-Smirnov Z		2.261	.682
Asymp. Sig. (2-tailed)		.000	.741

a Test distribution is Normal.
b Calculated from data.

4.4.1.4 I-695

2229 pairs of cellular vehicle speeds are obtained on I-695. The summary of deviation is shown in Table 4-3. The average cellular speed underestimates the average vehicular speed by 0.19 mile per hour and the average absolute error is 21.17%. More than 76.04% of records have an absolute error less than 20 percent. 17.55 percent of records have an absolute error between 20 to 50 percent. Neither type of speeds are normally distributed (Table 4-10). Non-parametric tests results (Table 4-11) indicate that these two groups are similar to each other.

Table 4-9
Test statistics

N(a)	91	N(b)	91		Cellular-Vehicular
Chi-Square	.275	Kendall's W(b)	.003	Z(c)	-.408
Df	1	Chi-Square	.275	Asymp. Sig. (2-tailed)	.684
Asymp. Sig.	.600	Df	1		
		Asymp. Sig.	.600		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

Table 4-10
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		2229	2229
Normal Parameters(a,b)	Mean	55.2400	55.4312
	Std. Deviation	10.88496	12.69871
Most Extreme Differences	Absolute	.125	.202
	Positive	.119	.108
	Negative	-.125	-.202
Kolmogorov-Smirnov Z		5.894	9.521
Asymp. Sig. (2-tailed)		.000	.000

a Test distribution is Normal.
b Calculated from data.

Table 4-11
Test statistics

N(a)	2229	N(b)	2229		Cellular-Vehicular
Chi-Square	1.669	Kendall's W(b)	.001	Z(c)	-1.946
Df	1	Chi-Square	1.669	Asymp. Sig. (2-tailed)	.052
Asymp. Sig.	.196	Df	1		
		Asymp. Sig.	.196		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

4.4.1.5 I-895

210 pairs of cellular vehicular speeds are collected on I-895. 77.14 percent of records have an absolute error less than 20% of the corresponding vehicular link speed. The average absolute error is 42%. The One Sample Kolmogrov-Smirnov test results show that neither type of speeds is normally distributed (Table 4-12). The statistical tests results shown in Table 4-13 confirm that these two groups of data are from the same population and these two groups of data are similar to each other.

4.4.1.6 US-1

430 pairs of cellular vehicle speeds are obtained on US-1 for this comparison. The average cellular speed overestimates the average vehicular speed by 23.59 miles per hour and the average absolute error is 151.21%. Only 23.49% of records have an absolute error less than 50 percent. 26.74% of records have an absolute error between 50 to 100 percent. 49.77% of records have an absolute error over 100 percent. Neither type of speeds is normally distributed (Table 4-14). The test results shown in Table 4-15 indicate that these two groups' speeds are not from the same population. One possible explanation is that arterial data used more default values since the volume of cell data was lower than that for the expressways. The use of historical average speeds as a default might have resulted in a different conclusion since the speed limit can be quite different than actual speed during periods of congestion.

Table 4-12
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		210	210
Normal Parameters(a,b)	Mean	55.4658	52.3720
	Std. Deviation	7.08693	12.88224
Most Extreme Differences	Absolute	.097	.241
	Positive	.097	.161
	Negative	-.071	-.241
Kolmogorov-Smirnov Z		1.402	3.498
Asymp. Sig. (2-tailed)		.039	.000

a Test distribution is Normal.
b Calculated from data.

Table 4-13
Test statistics

N(a)	210	N(b)	210		Cellular-Vehicular
Chi-Square	.076	Kendall's W(b)	.000	Z(c)	-.644
Df	1	Chi-Square	.076	Asymp. Sig. (2-tailed)	.520
Asymp. Sig.	.783	Df	1		
		Asymp. Sig.	.783		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

Table 4-14
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		430	430
Normal Parameters(a,b)	Mean	47.7289	24.1359
	Std. Deviation	7.03492	9.83582
Most Extreme Differences	Absolute	.501	.077
	Positive	.372	.077
	Negative	-.501	-.073
Kolmogorov-Smirnov Z		10.386	1.590
Asymp. Sig. (2-tailed)		.000	.013

a Test distribution is Normal.
b Calculated from data.

Table 4-15
Test statistics

N(a)	430	N(b)	430		Cellular-Vehicular
Chi-Square	368.38	Kendall's W(b)	.857	Z(c)	-17.574
Df	1	Chi-Square	368.38	Asymp. Sig. (2-tailed)	.000
Asymp. Sig.	.000	Df	1		
		Asymp. Sig.	.000		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

4.4.1.7 MD-40

519 pairs of cellular vehicle speeds are obtained. The average cellular speed overestimates the average vehicular speed by 14.37 miles per hour and the average percentage of absolute error is 96.70%. 21.77% of records have an absolute error less than 20 percent. 27.17% of records have an absolute error between 20 to 50 percent. 22.93% of records have an absolute error between 50 to 100 percent. The remaining 28.13% of records have an absolute error over 100 percent. The One Sample Kolmogrov-Smirnov test results show that neither type of speeds is normally distributed (Table 4-16). Non-parametric tests results confirm that these two groups' speeds are not from the same population. Table 4-17 shows that the cellular and vehicular speeds in this case are not from the same population.

4.4.1.8 MD-45

415 pairs of cellular vehicle speeds are obtained on MD-45. The average cellular speed overestimates the average vehicular speed by 23.14 miles per hour and the average absolute error is 167.02%. Only 4.34% of records have an absolute error less than 20 percent. 13.01% of records have an absolute error between 20 to 50 percent. 26.27% of records have an absolute error between 50 to 100 percent. More than 56.39% of records fall outside the 100% region.

Cellular speeds are not normally distributed, while vehicular speeds have normal distribution (Table 4-18). Non-parametric tests results (Table 4-19) indicate that these two groups' speeds are not similar to each other.

Table 4-16
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		519	519
Normal Parameters(a,b)	Mean	42.3165	27.9448
	Std. Deviation	14.24674	11.94743
Most Extreme Differences	Absolute	.303	.065
	Positive	.213	.065
	Negative	-.303	-.060
Kolmogorov-Smirnov Z		6.905	1.477
Asymp. Sig. (2-tailed)		.000	.026

a Test distribution is Normal.
b Calculated from data.

Table 4-17
Test statistics

N(a)	519	N(b)	519		Cellular-Vehicular
Chi-Square	211.100	Kendall's W(b)	.407	Z(c)	-15.792
Df	1	Chi-Square	211.100	Asymp. Sig. (2-tailed)	.000
Asymp. Sig.	.000	Df	1		
		Asymp. Sig.	.000		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

Table 4-18
Results of One Sample Kolmogrov-Smirnov Test

		Cellular Speed	Vehicular Speed
N		415	415
Normal Parameters(a,b)	Mean	43.5700	20.4348
	Std. Deviation	6.17516	8.63515
Most Extreme Differences	Absolute	.498	.064
	Positive	.379	.064
	Negative	-.498	-.041
Kolmogorov-Smirnov Z		10.144	1.303
Asymp. Sig. (2-tailed)		.000	.067

a Test distribution is Normal.
b Calculated from data.

Table 4-19
Test statistics

N(a)	415	N(b)	415		Cellular-Vehicular
Chi-Square	360.889	Kendall's W(b)	.870	Z(c)	-17.284
Df	1	Chi-Square	360.889	Asymp. Sig. (2-tailed)	.000
Asymp. Sig.	.000	df	1		
		Asymp. Sig.	.000		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

4.4.1.9 Martin Luther King Boulevard

133 pairs of cellular vehicle speeds are obtained on Martin Luther King Boulevard for link speed comparison. The average cellular speed overestimates the average vehicular speed by 17.38 miles per hour and the average absolute error is 163.67%. Only 9.02% of records have an absolute error less than 20 percent. 14.29% of records have an absolute error between 20 to 50 percent. 33.08% of records have an absolute error between 50 to 100 percent. More than 43.61% of records fall outside the 100% error region.

The One Sample Kolmogrov-Smirnov test results show that the cellular speeds are not normally distributed (Table 4-20). The statistical tests results (Table 4-21) show that these two groups' speeds are not from the same population, that is, these two groups of data are not similar to each other.

Table 4-20
Results of One Sample Kolmogrov-Smirnov test

		Cellular Speed	Vehicular Speed
N		133	133
Normal Parameters(a,b)	Mean	38.8349	21.4499
	Std. Deviation	17.88734	10.39717
Most Extreme Differences	Absolute	.267	.109
	Positive	.134	.109
	Negative	-.267	-.098
Kolmogorov-Smirnov Z		3.079	1.257
Asymp. Sig. (2-tailed)		.000	.085

a Test distribution is Normal.
b Calculated from data.

Table 4-21
Test statistics

N(a)	133	N(b)	133		Cellular-Vehicular
Chi-Square	56.910	Kendall's W(b)	.428	Z(c)	-7.677
Df	1	Chi-Square	56.910	Asymp. Sig. (2-tailed)	.000
Asymp. Sig.	.000	df	1		
		Asymp. Sig.	.000		

a Friedman Test
b Kendall's Coefficient of Concordance
c Wilcoxon Test

The statistical tests results confirm the similarity between cellular link speeds and vehicular link speeds on all freeway routes, except I-70. It should be noticed that relatively large error measures do not necessarily mean low similarity in data. A typical example is the data on I-95. Though the absolute percentage of error (63.54%) is much higher than that on I-70 (15.27%), the statistical tests results confirm the similarity of two types of link speeds on I-95, while reject the similarity on I-70.

On all arterial routes, the statistical test results indicate that cellular link speeds are not similar to vehicular link speeds.

4.4.2 Summary by Time of Day

The survey data spans from 6 am in the morning to 7 pm in the evening. The number of records in each hour ranges from 71 to 605. Table 4-22 shows the average speeds and average absolute errors during a day. The average absolute error ranges from 8.6 mile per hour (6 am to 7 am) to 13.7 mile per hour (1 pm to 2 pm). During morning peak and afternoon peak hours the data quality measures degrade.

Figure 4-5 illustrates the comparison of average link speeds by time of day. The two time intervals with lower average speeds correspond to morning peak hours and afternoon peak hours. Figure 4-6 closely illustrates the comparison of errors by time of day. It is should also be noted that beside the two peak hours, the error measures at noon time are extraordinary large.

A more detailed analysis of absolute error on each route by time of day is shown in Table 4-23. From this table, it is clear that the absolute errors on arterials are much larger than those on freeways. Most routes show a clear morning peak and afternoon peak (subject to data availability), I-395 shows a noon peak and an afternoon peak, and MD-40 shows a noon peak. Figure 4-7 illustrates the variation of average absolute errors on I-95, I-695, MD-40 and Martin Luther King Boulevard, since the data on these four routes spans most hours during a day.

Table 4-24 shows the count of records on each route during the day. There are 97 records during 1 PM to 2 PM and 105 records during 2 PM to 3 PM on MD 45 and the absolute error during those two hours are 23.78 mph and 23.22 mph respectively. However, when calculating the average percentage of absolute error, the value is 206.30% during 1 PM to 2 PM and 76.19% during 2 PM and 3PM. Part of the reason is explained in Figure 4-8. The 97 records on MD-45 with large absolute error make up to 25% of the records during 1 PM to 2 PM, while the 105 records make up to 20 % of the records during 2 PM to 3 PM.

Table 4-22
Summary of absolute error by time of day

Time	Vehicular Speed (mph)	Cellular Speed (mph)	Average Error (mph)	Avg. Abs. Error (mph)	Count
6:00-7:00	53.04523	58.63285	5.59	8.60	71
7:00-8:00	42.66473	52.09571	9.43	12.88	376
8:00-9:00	44.36707	49.84374	5.48	12.87	507
9:00-10:00	48.30601	54.76771	6.46	11.60	376
10:00-11:00	48.1949	56.1965	8.00	10.71	278
11:00-12:00	53.0143	57.97727	4.96	10.32	504
12:00-13:00	52.22896	56.38478	4.16	11.18	361
13:00-14:00	45.33864	52.0558	6.72	13.70	364
14:00-15:00	44.84126	53.35561	8.51	13.64	502
15:00-16:00	50.25374	51.85631	1.60	11.02	432
16:00-17:00	42.54214	47.49513	4.95	12.97	605
17:00-18:00	39.58631	45.93809	6.35	13.19	502
18:00-19:00	49.92598	53.28225	3.36	11.39	158

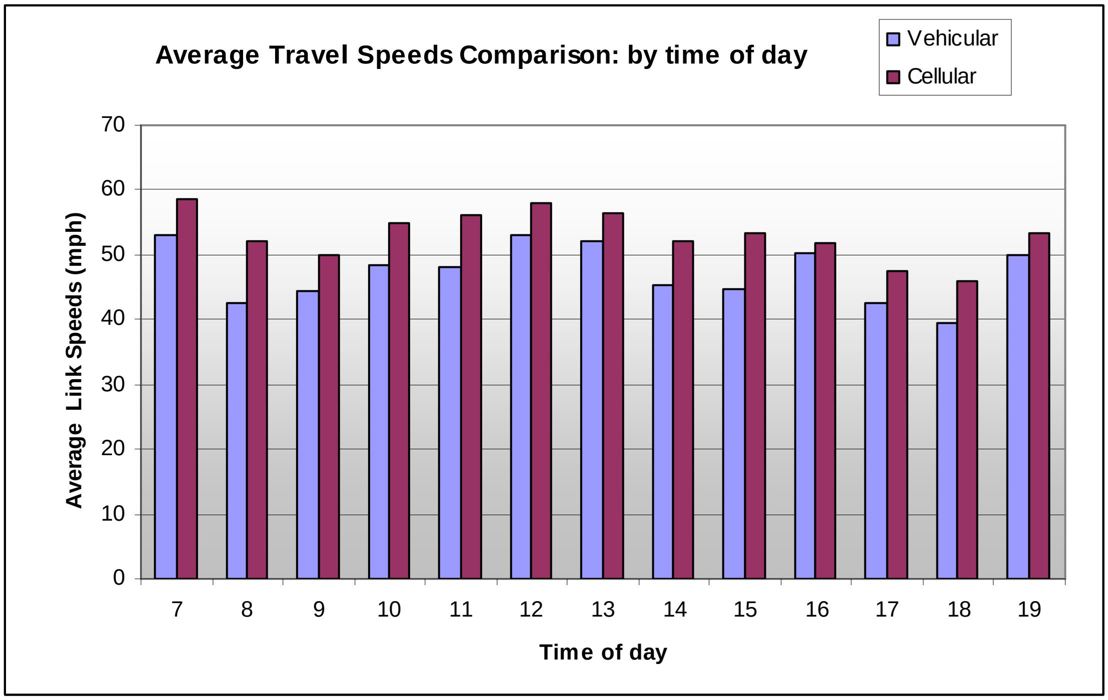

Figure 4-5
Comparison of average speeds: by time of day

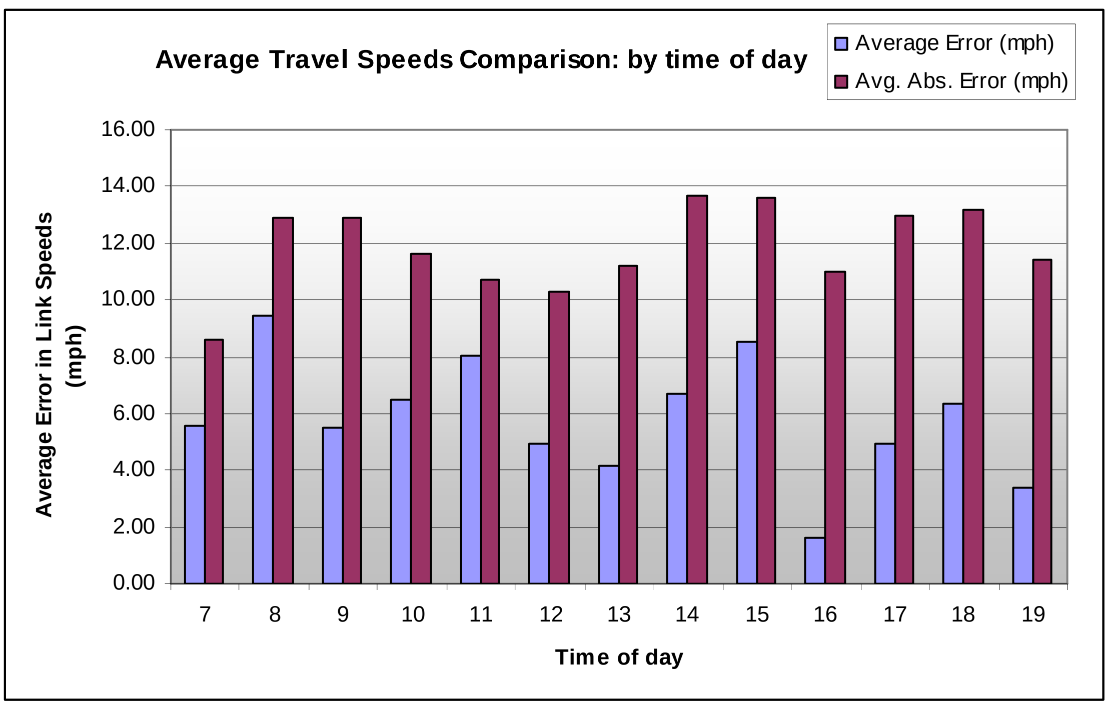

Figure 4-6
Comparison of average errors: by time of day

To further explore the reason why the average percentage of absolute errors during noon time is that large, we categorize the absolute errors into 5 groups: 0 to 5 mph, 5 to 10 mph, 10 to 20 mph, 20 to 50 mph and above 50 mph. It is observed that there are 8 very low vehicular speeds during 1 PM to 2 PM, as shown in Table 4-25. Since the percentage of absolute error is based on vehicular speeds, when the denominator is small and numerator is relatively large, some extremely large percentages are generated very easily. For example, when the vehicular speed is 4 mph, and the cellular speed is 50 mph, the percentage of error is 1150%! These outliers dramatically degrade the average value of errors. As shown in Figure 4-9, there are more than 10% of records during 1 PM to 2 PM that are lower than 5 miles per hour. These outliers greatly increase the average percentage of absolute error.

Therefore, one must be cognizant of the impact of these outlier points when examining the results presented in this report.

Table 4-23
Absolute error by time of day on different routes

Time	I-70	I-395	I-695	I-895	I-95	US1	MD40	MD45	MLKB
6:00-7:00			10.08		7.39				
7:00-8:00		7.84	8.73		10.19	22.69	18.47		26.04
8:00-9:00		9.98	8.70		12.13	23.87	16.53		23.36
9:00-10:00		8.27	7.05		10.47	25.02	17.79		26.73
10:00-11:00			5.55		4.11	25.90	21.49		
11:00-12:00		3.60	4.57		10.44	24.96	20.39		24.53
12:00-13:00		14.33	5.62		9.68	23.13	23.54	23.78	24.23
13:00-14:00		18.22	6.62		8.73		18.03	23.22	22.99
14:00-15:00			9.53	12.26	7.08		18.06	23.08	
15:00-16:00	6.21	9.32	9.32	18.62	8.13		16.33	25.13	15.28
16:00-17:00	9.35	15.88	9.32	10.70	9.72		16.72	23.45	18.61
17:00-18:00	6.89	14.32	10.40	7.88	11.41		14.23	25.58	13.36
18:00-19:00		20.21	9.48	8.07	12.22		13.78		

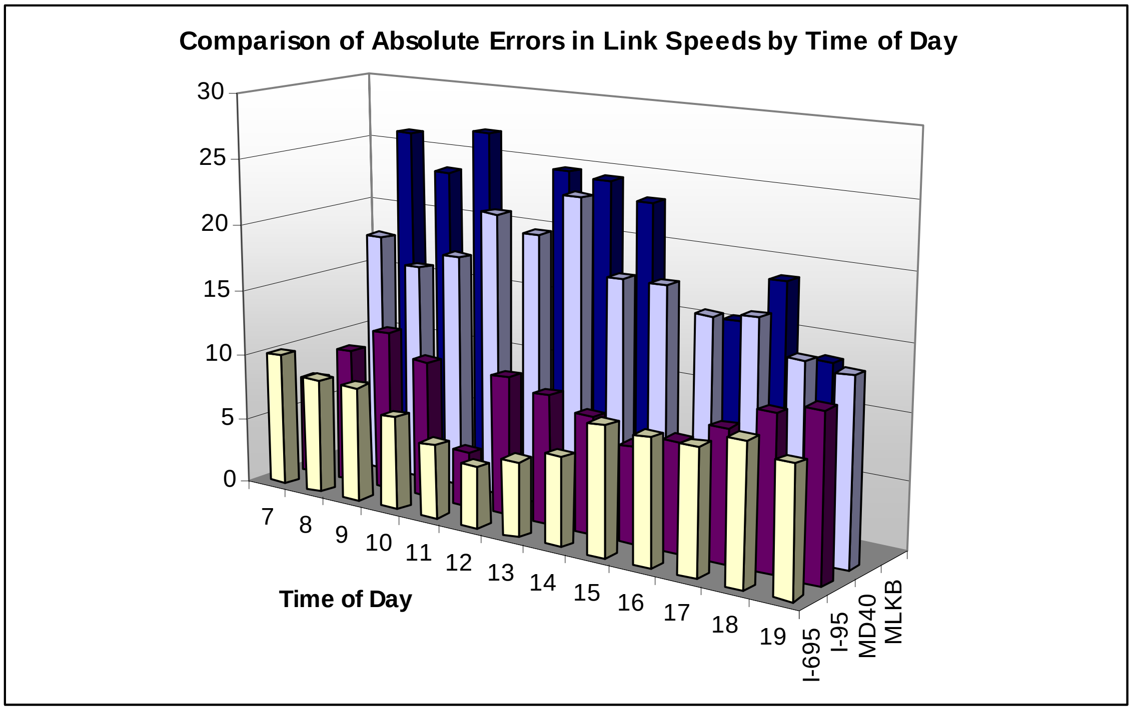

Figure 4-7
Comparison of average errors in link speeds: by time of day and route

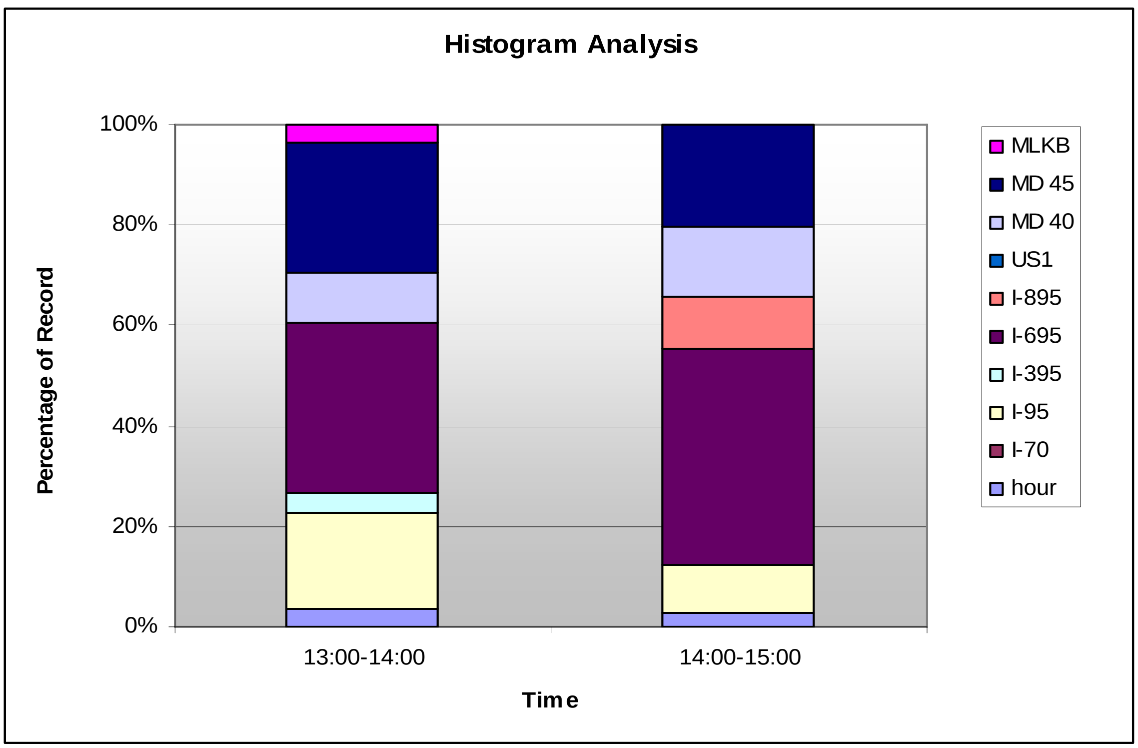

Figure 4-8
Histogram analysis of data records at noon

Table 4-24
Summary of absolute error by time of day

Time	I-70	I-95	I-395	I-695	I-895	US1	MD40	MD45	MLKB
6:00-7:00		39		32					
7:00-8:00		85	15	190		79	38		21
8:00-9:00		92	18	281		95	52		28
9:00-10:00		95	9	207		85	22		14
10:00-11:00		43		170		79	33		
11:00-12:00		230	2	197		64	47		4
12:00-13:00		215	12	85		28	4	9	24
13:00-14:00		73	14	129			37	97	14
14:00-15:00		50		221	54		72	105	
15:00-16:00	7	86	3	249	3		52	28	4
16:00-17:00	29	68	10	231	85		78	94	10
17:00-18:00	22	68	5	178	64		69	82	14
18:00-19:00		77	3	59	4		15		

Table 4-25
Histogram analysis of average speeds by time of day

Time	0~5 mph	5~10 mph	10~20 mph	20~50 mph	>50 mph
6:00-7:00		1	7	30	
7:00-8:00	4	1	10	35	2
8:00-9:00	1		6	15	
9:00-10:00		3	5	25	
10:00-11:00		1	8	35	3
11:00-12:00			2	2	
12:00-13:00			10	25	2
13:00-14:00		8	21	42	1
14:00-15:00		2	14	35	1
15:00-16:00	1	3	23	50	1
16:00-17:00	1	4	20	43	1
17:00-18:00			6	8	1
18:00-19:00	3	59		4	77

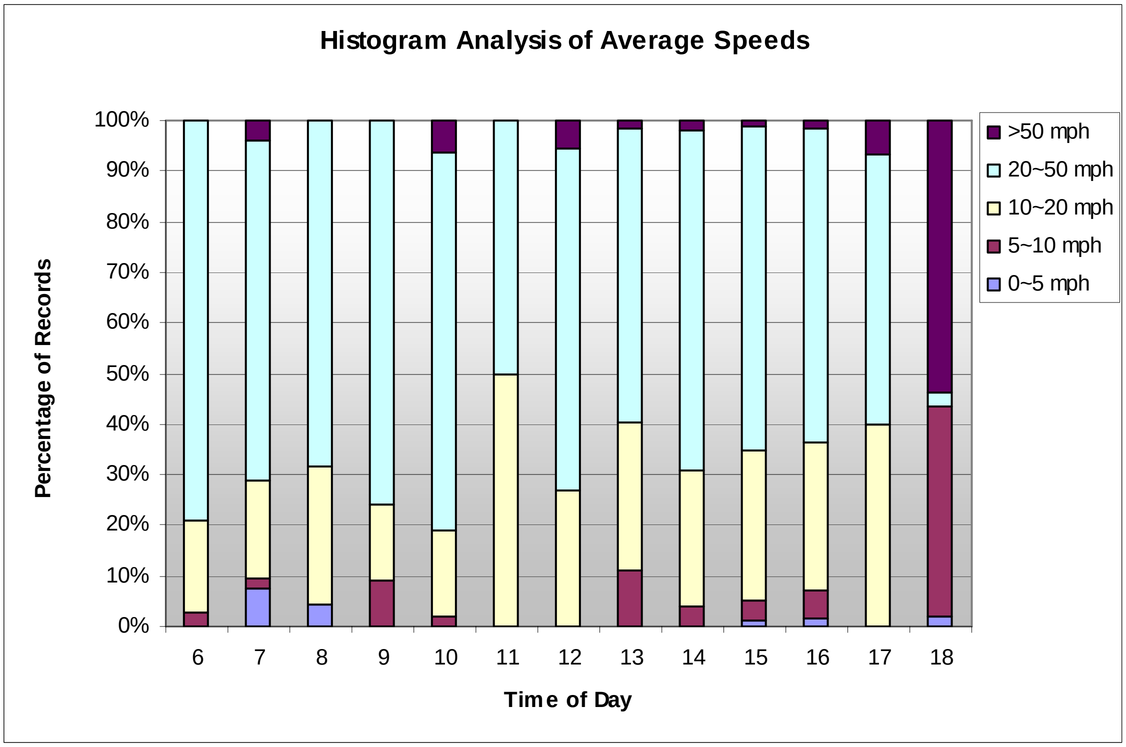

Figure 4-9
Histogram analysis of average speed: by time of day

4.4.3 Summary by Level of Service (LOS)

The level of service is reported by the probe vehicle drivers according to their personal experience in traffic. Six levels are available from Level A to F, with Level A being the best and Level F being the worst.

Figure 4-10 shows both the percentage of records in each LOS for all 9 routes, with the last column showing the overall LOS distribution. The tabulated results are presented in Table 4-26. In general, interstate freeways have better LOS distributions than the arterials, i.e., more data points fall into high levels for interstate freeways than the arterials. If we look at interstates I-95, I-695 and I-895, we can observe that over 80% of the data fall into Levels A and B. However, for arterials US-1 and MD-45, almost all data are in Levels D, E, and F. This indicates that LOSs on arterials are worse than those on Interstates. Overall, more than 50.77% of the data are in Level A, 8.16% in Level B, 4.45% in Level C, 6.14% in Level D, 15.15% in Level E, and 15.32% in Level F.

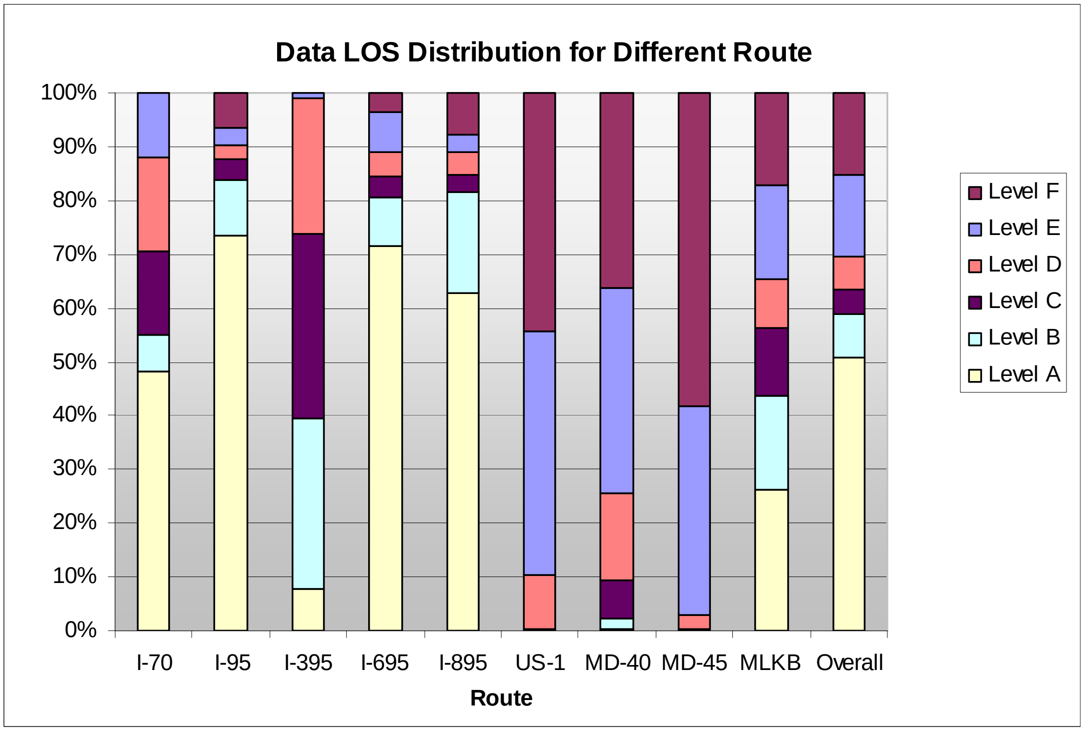

Figure 4-10
Data LOS distribution for different routes

Considering data for all routes as a whole, we can sort them into different groups based on their LOS. For example, all data labeled Level A are grouped together for speed comparison, and the same is done for the other five levels. Table 4-27 presents the distribution of absolute errors in link speeds for different Levels of Service.

It is obvious that with the decrease of LOS (from Level A to F), fewer data have small absolute error in percentage. For example, more than 85% of the data have absolute error between 0 and 20% for Level A. While the corresponding value for Level B drops to 63%, 46% for Level C, 40% for Level D, and 12% for Level E. Only 4% of the data are within 0-20% of absolute error for Level F. It is observed that over 82% of the data have a percentage over 100% for Level F. Figure 4-11 plots the absolute error for different LOS.

Besides absolute error, absolute speed difference can also be calculated for different Levels of Service to compare its distribution. Table 4-28 summarizes the distribution of absolute difference for each level. The absolute speed differences are categorized into four groups: 0-5 mph, 5-10 mph, 10-20 mph, and >20 mph. The plotted results are shown in Figure 4-12.

Table 4-26
Data LOS distribution for different routes

Route	Level A	Level B	Level C	Level D	Level E	Level F
I-70	48.28%	6.90%	15.52%	17.24%	12.07%	0.00%
I-95	73.55%	10.16%	3.85%	2.78%	3.28%	6.39%
I-395	7.69%	31.87%	34.07%	25.27%	1.10%	0.00%
I-695	71.47%	9.11%	3.90%	4.40%	7.72%	3.41%
I-895	62.86%	18.57%	3.33%	4.29%	3.33%	7.62%
US-1	0.00%	0.00%	0.23%	10.23%	45.35%	44.19%
MD-40	0.19%	2.12%	6.94%	16.38%	38.15%	36.22%
MD-45	0.00%	0.00%	0.24%	2.65%	38.80%	58.31%
MLKB	26.32%	17.29%	12.78%	9.02%	17.29%	17.29%
Overall	50.77%	8.16%	4.45%	6.14%	15.15%	15.32%

Table 4-27
Absolute error for different LOS

Abs. Error	Level A	Level B	Level C	Level D	Level E	Level F
0-10%	0.573	0.3843	0.2638	0.1631	0.066	0.0209
10%-20%	0.2837	0.2523	0.2043	0.2369	0.061	0.0234
20%-50%	0.1229	0.2523	0.3234	0.3815	0.3039	0.0517
50%-100%	0.0167	0.0463	0.0213	0.12	0.4259	0.0751
>100%	0.0037	0.0648	0.1872	0.0985	0.1432	0.8288

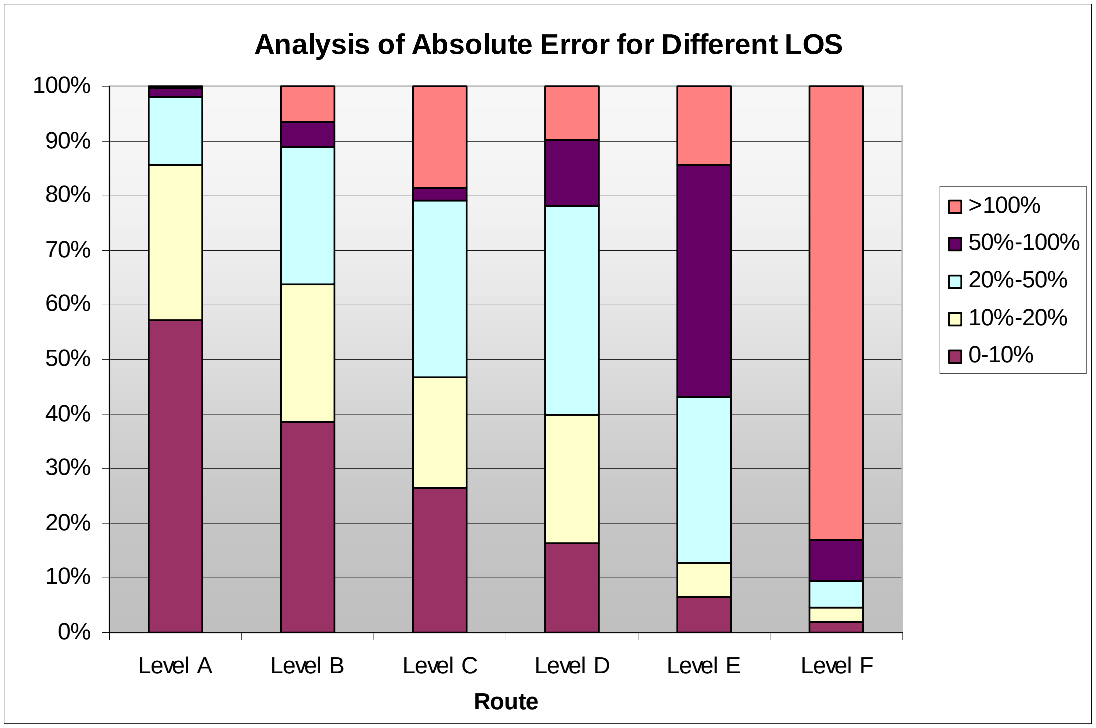

Figure 4-11
Absolute error for different LOS

Table 4-28
Absolute speed difference for different LOS

Abs. Diff	Level A	Level B	Level C	Level D	Level E	Level F
0-5 mph	0.4939	0.3796	0.3362	0.2308	0.1046	0.0751
5-10 mph	0.2941	0.2847	0.234	0.3169	0.127	0.048
10-20 mph	0.1556	0.2708	0.3234	0.3138	0.4321	0.0727
>20 mph	0.0564	0.0648	0.1064	0.1385	0.3362	0.8042

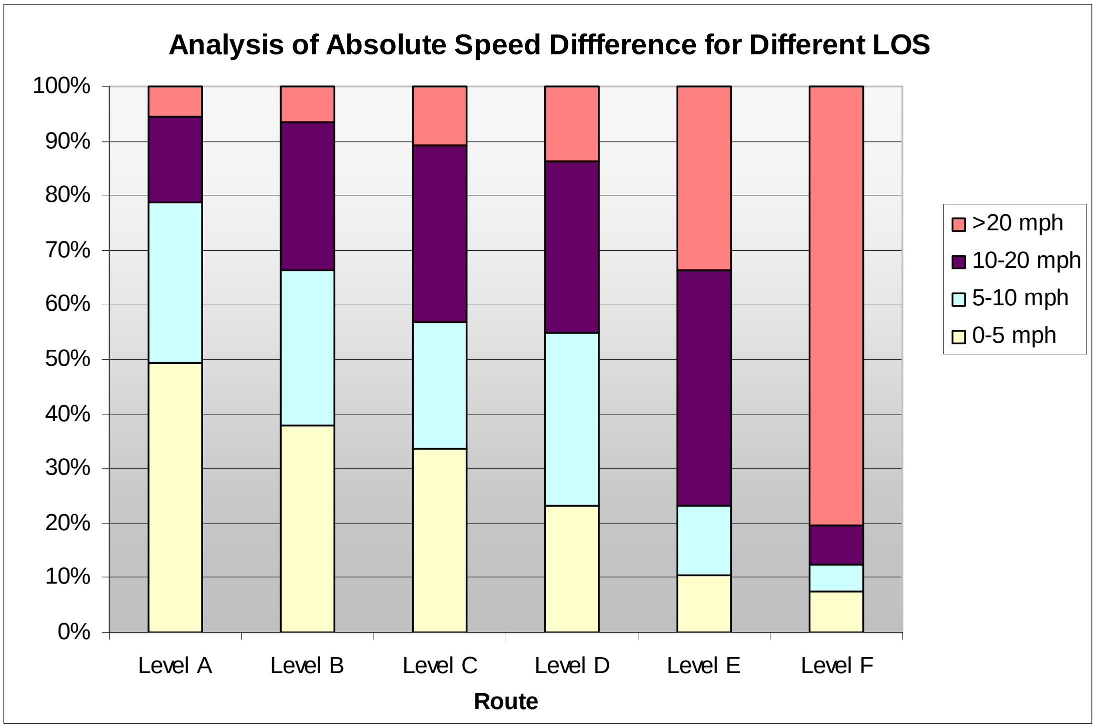

Figure 4-12
Absolute speed difference for different LOS

The same trend as the one discussed before for the absolute error is observed in the absolute speed difference. More and more data tend to have large speed difference when the Level of Service is decreased. For Level A, over 78% of the records have an absolute speed difference between 0 to 10 mph, while only 7.9% of the records are in the same group for Level F.

4.5 Summary

In this chapter, we examined the link traffic information by evaluating link travel speeds. The reason to evaluate link travel speeds instead of link travel times is due to the link length discrepancy. The link speed analyses show the average absolute errors on freeways are about 10 miles per hour and those on arterials are around 20 miles per hour. The percentile analysis of absolute error shows the same trend. The statistical analyses confirm that cellular speeds and vehicular speeds are similar to each other on freeways (except on I-70), and also indicate that these two sets of data are not similar on all surveyed arterials.

During morning peak hours and afternoon peak hours, the data quality measures degrade. The unexpected large average percentage of error at noon was explained by the analysis of outlier impact, which is caused by some extremely low vehicular link speeds. However, in the calculation of absolute errors of link speeds, when there are a large amount of records, few outliers (e.g. 50 mile per hour difference) will not have a major impact on the average absolute error; when the number of records is small, the outliers represent a significant portion of the data and they cannot be neglected. For this reason we did not remove these outliers from the calculations of data quality measures. Because of this outlier impact, the data quality analyses are mainly based on absolute errors rather than percentage of absolute errors.

The link speed evaluation with respect to the level of service shows that the quality measures have significant differences in low LOS and high LOS. The quality measures dramatically degrade when the LOS drops.

5 Path Traffic Information Evaluation

5.1 Path Length Discrepancy

Each survey run is deemed as a path, unless there is only one link that can be matched in a run. In the latter situation, the path data quality evaluation is the same as the link data evaluation. Travel time is used for path data quality evaluation. Compared to link lengths, there is less length discrepancy for paths. The summary of path length analysis is shown in Table 5-1. 65.45% of paths have a path length discrepancy less than 10%. Those records with small length discrepancies are used in path travel time comparison. When the path discrepancy is large, average travel speed on a path should be used for data quality evaluation.

Table 5-1
Histogram of Path Lengths

% of Abs. Error			Frequency		Probability Density	
[0, 10%)			322		65.45%	
[10%, 20%)			94		19.11%	
[20%, 50%)			72		14.63%	
[50%, 100%)			4		0.81%	
>=100%			0		0%	
	Cell (mile)	Vehicle (mile)	Error (mile)	% Error	Abs. Error (mile)	Abs. % Error
Avg.	8.97	9.45	0.48	5.35	0.61	6.74

5.2 Data Processing

The following procedure is performed for data preparation in order to compare the travel times on paths:

Step 0: Combine and sort segmented probe vehicle data for each run;

Sstep1: Catch exceptions such as nodes without matching node numbers or potential errors in matched node numbers in probe vehicle data;
Step 2: Find the start time, end time, start node and end node of a survey runs. The start node is the second node that can be matched in a survey run, since the TMC link ID is the destination node of a link;
Step 3: Calculate vehicular path travel times by subtracting the start time from the end time.
Step 4: Query road network data to define the direction for each link. Transfer date and time to the same format as in cellular probe data;
Step 5: Construct a complete path based on the start node and end node information;
Step 6: Call PathTimeCalculator to calculate cellular path travel time.

The PathTimeCalculator procedure is presented as below:

Step 0 Initialization:
CurrentNode=StartNode
Direction=Cell_Direction
TimeParameter=StartTime of Path

Step 1 Run query:
Get two travel times in two adjacent 5-minute intervals.
Denote TimeStamp(Minus) = the one no greater than the TimeParameter,
TimeStamp(Plus) = the one greater than the TimeParameter

Step 2 Travel Time Calculations:

```
  IF TimeStamp(Plus)-TimeParameter>= TravelTime(Minus)
  THEN  NextTimeParameter= TimeParameter+TravelTime(Minus)
  IF NextNode!=Null
CurrentNode= NextNode
Goto Step 1
  Else End
  ELSE NextTimeParameter= TimeStamp(Plus)+(Length - (TimeStamp(Plus) -
TimeParameter)* Speed(Minus))/Speed(Plus)
   IF NextNode!= Null
              CurrentNode=NextNode
Goto Step 1
Else End
```

An example of data processing is demonstrated in Table 5-2 and 5-3. Table 5-2 shows the probe vehicular data of the survey run that starts from 7:28:01am. The first node that can be matched is 4424. The survey run ended at 7:47:56 am, at node 4442. The entire path travel time in this case is 1101 second. In Table 5-3, a complete path is constructed. Instead of the “Tunnel”, a valid intermediate node 4432 (colored in orange) is found from the network file, which links node 4431 and 4672. The green cells are those TMC links which need travel time interpolation. The calculated path travel time is 1059 seconds. The error is smaller than 4 percent.

5.3 Data Quality Measures

Similar to the measures used in link speed evaluations, the measures of travel time are based on the data obtained by probe vehicle data as well. The actual travel time is deemed as "Ground Truth" and is calculated from the GPS recordings and then compared with the corresponding travel time reported from the ITIS cellular traffic system (CFVD™ Time). Again, it should be noted that due to the limited number of probe vehicles the collected vehicular travel time data may vary from the "Ground Truth". However, the vehicular travel time is used as "Ground Truth" in the calculation of data quality measures.

The system error calculated is calculated as follows:
Error = (CFVD™ Time -"Ground Truth") (5-1)
Percentage of Error = (CFVD™ Time- "Ground Truth")/ "Ground Truth" (5-2)

The same 4 measures (Average Error, Average Percentage of Error, Average Absolute Error and Percentage of Average Absolute Error) are calculated for the entire data set and for specific routes. The statistical tests are calculated for each route.

5.4 Summary of Statistical Tests

The summary of cellular travel time and vehicular travel time deviations is shown in Table 5-4. It is notable that the average percentage of absolute error on freeways (I-70, I-95, I-395, I-695, I-895) is much smaller than that on arterials (US-1, MD-40, MD-45, Martin Luther King Boulevard). Wilcoxon test result indicates that the means of error percentages of these two types of roads have significant difference. While on freeway the average absolute percentage of error is around 10%, those on arterials often exceed 50% of actual vehicular travel time. The 50 percentile absolute errors are in a range of 5.2% to 9.3% on freeways (except I-395), while the corresponding values on arterials range from 34.7% to 69.4%. The 85 percentile absolute errors give a range of the absolute errors for most records. Here, only the percentile analysis on absolute error in percentage is provided since the average travel time on each route varies.

Table 5-2
Probe vehicle record

TIMESTAMP	Path Time	Node	length(m)	Speed(kph)	TravelTime_Link
7:28:01	0	Enter	0.000	0.000	0
7:29:35	94	4424	2751.978	104.768	94
7:30:44	163	4425	1899.026	100.101	68
7:31:58	237	4426	2059.960	100.101	74
7:33:08	307	4427	1915.119	98.009	71
7:33:15	314	4428	193.121	107.021	7
7:33:46	345	4429	933.420	108.631	31
7:34:02	361	4430	498.897	106.217	17
7:34:57	416	4431	1528.877	100.423	55
7:35:13	432	Tunnel	434.523	95.434	16
7:36:49	528	Tunnel	2156.521	81.111	96
7:37:42	581	4672	2945.100	64.171	165
7:37:57	596	4433	80.467	20.600	15
7:38:38	637	4434	1029.980	90.445	41
7:39:09	668	4435	917.326	107.021	31
7:39:48	707	4436	997.793	91.894	39
7:40:12	731	4437	579.364	87.387	24
7:41:08	787	4438	1689.811	107.021	57
7:41:30	809	4439	692.018	111.849	22
7:42:10	849	4440	1239.195	113.459	39
7:44:37	996	4441	4473.976	109.114	148
7:47:56	1195	4442	5343.022	96.561	199

Table 5-3
Calculation by cellular data

Cell.TIMESTAMP	TMCLinkID	Direction	Arrival Time	journeytime(secs)	Cell.length(m)
	4424	1	7:29:35		
7:25:00-7:30:00 AM	4425	1	7:30:38	63.23	1803
7:30:00	4426	1	7:31:39	61.2	1692
7:30:00	4427	1	7:32:25	45.75	1204
7:30:00	4428	1	7:33:04	39.29	1034
7:30:00	4429	1	7:33:10	5.05	133
7:30:00	4430	1	7:33:32	22.76	599
7:30:00	4431	1	7:34:25	53.01	1395
7:30:00 - 7:35:00 AM	4432	1	7:36:07	101.35	2667
7:35:00	4672	1	7:36:26	19.46	512
7:35:00	4433	1	7:36:48	22.19	584
7:35:00	4434	1	7:37:10	21.95	553
7:35:00	4435	1	7:38:06	55.95	1285
7:35:00	4436	1	7:38:38	32.28	888
7:35:00-7:40:00 AM	4437	1	7:39:15	36.98	1017
7:40:00	4438	1	7:39:54	38.24	1047
7:40:00	4439	1	7:40:26	32.09	888
7:40:00	4440	1	7:41:13	47.12	1337
7:40:00	4441	1	7:44:11	178.59	5261
7:40:00-7:45:00 AM	4442	1	7:47:14	182.81	5139

Figure 5-1 shows the comparison of average cellular travel time and average vehicular travel time on each route. It is noticed that the difference between these two types of travel times is much smaller on freeways than that on arterials. Figure 5-2 illustrated the comparison of the average error and the average absolute error on each route. On most arterials, the average error is very close to the average absolute error. This indicates that the cellular travel times consistently underestimated the vehicular travel times.

The histogram analysis of absolute errors shows the same trend as in the comparison of average errors (Figure 5-3). On freeways, more than 75% of records have absolute error less than 20% of actual vehicular travel time, while on arterials, the records in this range drops below 20%.

Table 5-4
Summary of travel time deviations

Route Name	Avg. Vehicular Travel Time (sec)	Avg. Cellular Travel Time (sec)	Avg. Abs. Error (sec)	Avg. % of Abs. Error (%)	50%kv Abs. Error (%)	85%kv Abs. Error (%)
I-70	550.31	539.17	29.56	6.81	5.21	9.32
I-95	1065.39	1007.17	105.53	9.78	8.73	16.90
I-395	55.35	48.99	6.80	11.59	23.75	31.60
I-695	1009.61	955.40	139.40	11.06	9.29	22.06
I-895	787.50	685.95	106.75	12.63	8.32	26.70
US-1a	885.78	501.68	384.10	39.99	45.44	51.92
US-1b	1079.95	323.43	756.52	69.41	70.24	73.85
MD-40	974.05	669.45	325.56	33.54	34.65	47.00
MD-45	955.52	381.10	574.42	58.56	60.75	65.16
MLKB	239.11	186.78	109.81	47.55	31.13	63.51

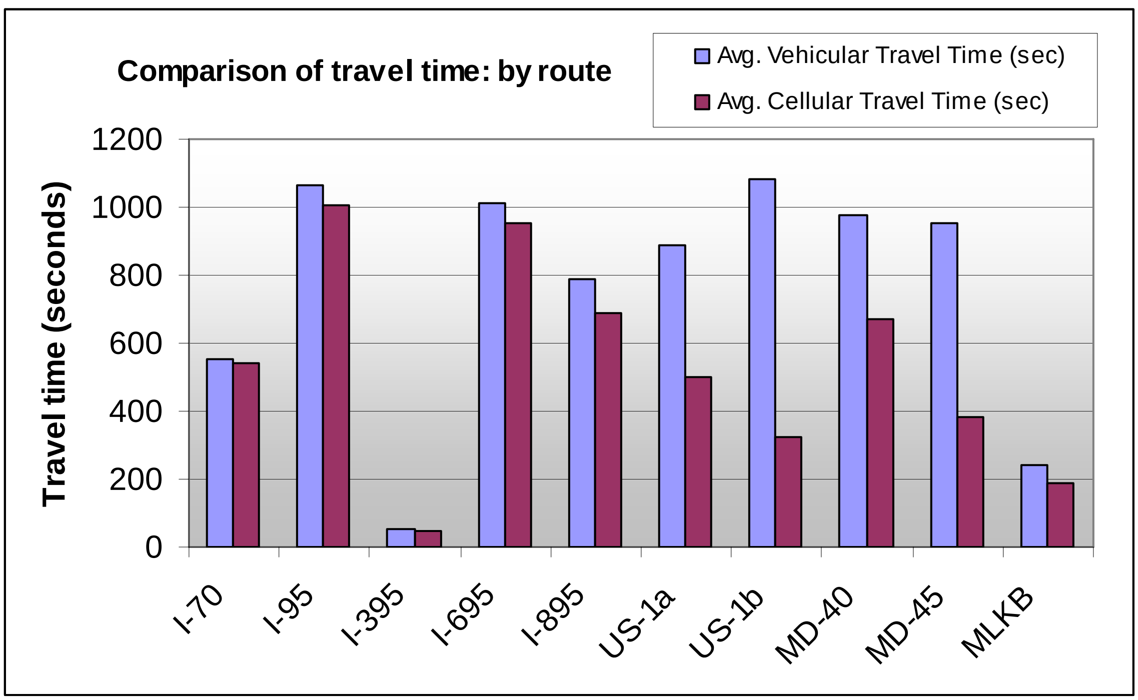

Figure 5-1
Comparison of average travel times: by route

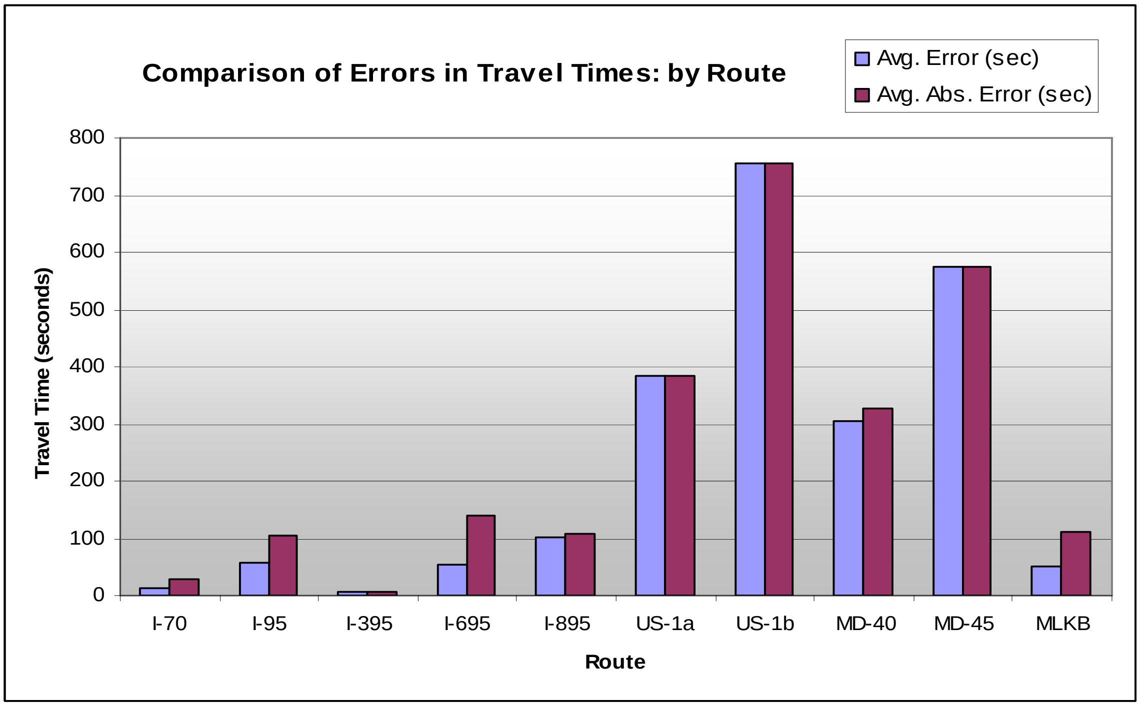

Figure 5-2
Comparison of average errors: by route

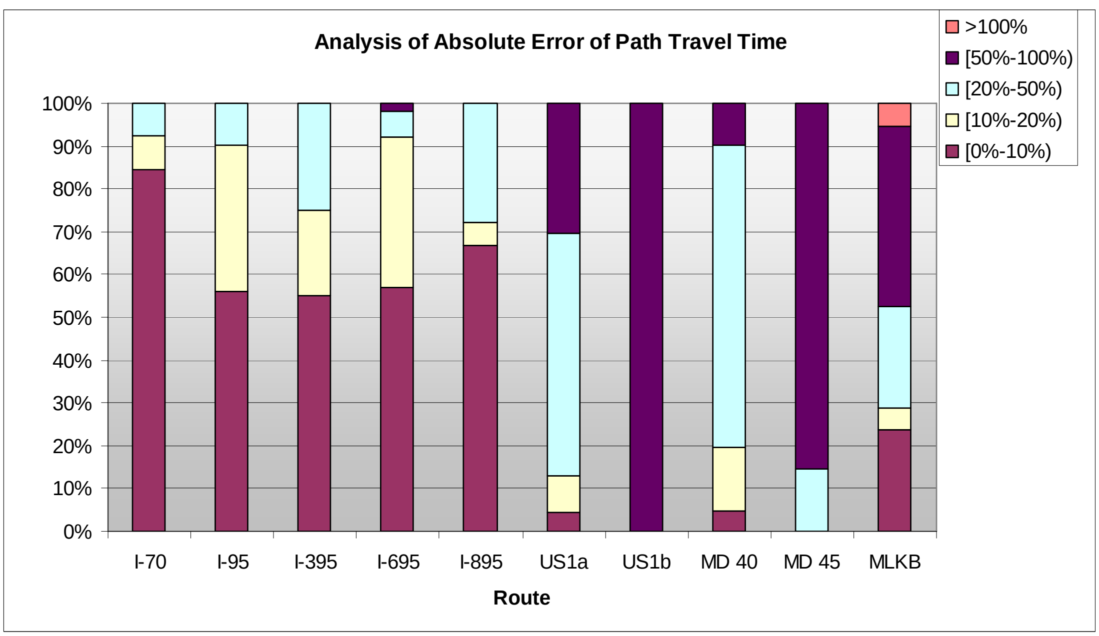

Figure 5-3
Analysis of percentage of absolute error

Similar to section 4.4.1., the results of travel time evaluation will be presented for routes. The error analysis and statistical test results on each route will be presented in the following sections.

5.4.1 Travel Time for I-70

Before analyzing the difference between cellular and vehicular path travel times, the path lengths were compared for.13 runs on I-70. 11 of 13 runs are about 10.5 miles long, and the other 2 runs are 4.16 and 0.41 miles long respectively. To prevent the potential impact of length discrepancy on travel time evaluation, two scenarios were analyzed. In scenario A, all the runs were tested together; in scenario B, the 11 runs longer than 10 miles were analyzed as a group.

The length discrepancy analysis shows these two scenarios have similar average absolute and average percentage of errors (Table 5-5). 11 runs are about 13% longer than the calculated path length, 1 run is between 10% to 20% and another run is between 20% to 50% of the calculated path length. The errors in both scenarios are very small. The average absolute errors in both scenarios are around 5%, which satisfy the recommended value (20%) discussed in Chapter 2. Considering the fact that the actual vehicular travel distance is 13% longer than the cellular path length; the path speed is also reported.

Table 5-5
Length discrepancy and error analysis on I-70

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (a) (mile)	9.18	7.91	1.27	13.80
Path Length (b) (mile)	10.43	9.05	1.38	13.24
Path Travel Time (a) (sec)	550.31	539.17	29.56	6.81
Path Travel Time (b) (sec)	626.27	618.04	29.34	4.79
Path Travel Speed (a) (mph)	59.37	52.85	6.72	11.06
Path Travel Speed (b) (mph)	60.32	53.04	7.28	11.95

The one sample Kolmogorov-Smirnov test shows the distribution of cellular travel time follows a normal distribution, while the distribution of probe vehicle travel time is significantly different from a normal distribution. Three non-parametric tests were performed to test the similarity of these two distributions. The test results show that these two groups of variables are from the same population. The test results are shown in Table 5-6.

Table 5-6
Statistical test results on I-70

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		13	13
Normal Parameters(a,b)	Mean	550.3077	539.1665
	Std. Deviation	195.38909	199.98792
Most Extreme Differences	Absolute	.392	.354
	Positive	.193	.215
	Negative	-.392	-.354
Kolmogorov-Smirnov Z		1.415	1.278
Asymp. Sig. (2-tailed)		.037	.076
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-1.153	
Asymp. Sig. (2-tailed)		.249	
N (a)		13	
Chi-Square		.692	
Df		1	
Asymp. Sig.		.405	
N (b)		13	
Kendall's W(b)		.053	
Chi-Square		.692	
Df		1	
Asymp. Sig.		.405	

a Friedman Test
b Kendall's Coefficient of Concordance

5.4.2 I-95

On I-95, 82 survey runs were analyzed. 73 of 81 are more than 10 miles long, 7 of them are around 5 miles long and the other 2 runs are 3.26 and 1.18 miles long respectively. Similar to the analysis on I-70, two scenarios were analyzed. In scenarios A, all the runs were tested together; in scenario B, the 72 runs longer than 10 miles were analyzed as a group.

The length discrepancy analysis shows these two scenarios have very small average percentage of error (less than 1 %). The errors in both scenarios are smaller than 10%, which satisfy the recommended value discussed in Chapter 2. The histogram analysis shows that 57.14% of records have absolute errors smaller than 10%, 33.33% of records have errors between 10% to 20% and 9.52 percent of records have errors between 20% to 50%.

The one sample Kolmogorov-Smirnov test results indicate that both travel time data show significant difference with the normal distribution. Three non-parametric tests results also show that there is significant difference (p=0.05) between the distributions of these two travel time data source. Table 5-7 shows the length discrepancy analysis of these two scenarios. Statistical results are presented in Table 5-8

Table 5-7
Length discrepancy and error analysis on I-95

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (a) (mile)	15.65	15.66	0.13	0.85
Path Length (b) (mile)	16.82	16.88	0.11	0.67
Path Travel Time (a) (sec)	1065.39	1007.17	105.53	9.78
Path Travel Time (b) (sec)	1148.65	1086.88	109.84	9.73

5.4.3 I-395

39 survey runs were collected on I-395. However, all of these survey runs are shorter than 1 mile. 20 out of 39 records have very small length discrepancies, while the other records have large discrepancies. Here we only present the result of the analysis for the group with small length discrepancy. When length discrepancy is large, the comparison of path travel time is less meaningful. The link speed comparison on I-395 was discussed in Section 4.4.1.

As shown in Table 5-9, the average percentage of absolute length discrepancy is 1.61%, which indicates the length discrepancy impact can be ignored in the path travel time evaluation. The average error is 11.59% of the vehicular path travel time. The histogram analysis shows that 55% of records have absolute errors smaller than 10%, 20% of records have errors between 10 to 20% and another 25 percent of records have errors betweens 20% to 50%.

On I-395, the probe vehicle travel times follow a normal distribution, while the cellular travel times do not. The Wilcoxon test result shows there is no significant difference between the means of these two variables but the Friedman test and Kendall's test show that these two groups of data are not from the same population (Table 5-10).

Table 5-8
Statistical test results on I-95

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		82	82
Normal Parameters(a,b)	Mean	1065.3902	539.1665
	Std. Deviation	365.42344	199.98792
Most Extreme Differences	Absolute	.216	.354
	Positive	.139	.215
	Negative	-.216	-.354
Kolmogorov-Smirnov Z		1.960	2.564
Asymp. Sig. (2-tailed)		.001	.000
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-3.067	
Asymp. Sig. (2-tailed)		.002	
N (a)		82	
Chi-Square		5.902	
Df		1	
Asymp. Sig.		.015	
N (b)		82	
Kendall's W (b)		.072	
Chi-Square		5.902	
Df		1	
Asymp. Sig.		.015	

a Friedman Test
b Kendall's Coefficient of Concordance

Table 5-9
Length discrepancy and error analysis on I-395

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	0.69	0.69	0.01	1.61
Path Travel Time (sec)	55.35	48.99	6.80	11.59

Table 5-10
Statistical test results on I-395

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		39	39
Normal Parameters(a,b)	Mean	53.4615	52.9785
	Std. Deviation	6.37366	31.75246
Most Extreme Differences	Absolute	.130	.350
	Positive	.119	.350
	Negative	-.130	-.187
Kolmogorov-Smirnov Z		.810	2.189
Asymp. Sig. (2-tailed)		.528	.000
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-2.763	
Asymp. Sig. (2-tailed)		.006	
N (a)		39	
Chi-Square		11.308	
Df		1	
Asymp. Sig.		.001	
N (b)		39	
Kendall's W(b)		.290	
Chi-Square		11.308	
Df		1	
Asymp. Sig.		.001	

a Friedman Test
b Kendall's Coefficient of Concordance

5.4.4 I-695

On I-695, 185 survey runs were analyzed. 162 runs are more than 10 miles long, 22 of them are around 5 miles long and only 1 run is 0.31 miles long. Since more than half of the paths have a length discrepancy more than 10%, we sorted the records according to the length discrepancy and selected those runs with length discrepancies less than 5% (51 runs) for path travel time analysis. After this pre-processing, the average absolute length discrepancy is less than 1 % (as shown in Table 5-11). The average absolute error in path travel time is 11.1%. The histogram analysis shows that 29 out of 51 records have absolute errors smaller than 10%, 18 out of 51 of records have errors between 10% to 20%, 3 runs have errors between 20% to 50%, and 1 run has an error of 72.7% of the probe vehicle path travel time.

The one sample K-S test results (Table 5-12) show the cellular travel time follows a normal distribution while the probe vehicle travel times do not. All three nonparametric tests show that these two groups of data have the same distribution.

Table 5-11
Length discrepancy and error analysis on I-695

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	13.78	13.89	0.11	0.96
Path Travel Time (sec)	1009.61	955.40	139.40	11.06

Table 5-12
Statistical test results on I-695

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		184	184
Normal Parameters(a,b)	Mean	919.5054	879.2720
	Std. Deviation	377.79874	291.07461
Most Extreme Differences	Absolute	.109	.076
	Positive	.109	.062
	Negative	-.080	-.076
Kolmogorov-Smirnov Z		1.485	1.032
Asymp. Sig. (2-tailed)		.024	.237
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-1.143	
Asymp. Sig. (2-tailed)		.253	
N(a)		184	
Chi-Square		.087	
Df		1	
Asymp. Sig.		.768	
N(b)		184	
Kendall's W(b)		.000	
Chi-Square		.087	
Df		1	
Asymp. Sig.		.768	

a Friedman Test
b Kendall's Coefficient of Concordance

5.4.5 I-895

18 survey runs were collected on this segment. 15 runs are around 11 miles long and 3 of them are around 7 miles long. The average absolute length discrepancy is 3.17% as shown is Table 5-13. The average absolute error in path travel time is 12.63%. The histogram analysis shows that 12 out of 18 records have absolute errors smaller than 10%, 1 out of 18 of records has an error between 10% to 20% and 5 runs have errors between 20% to 50% of the probe vehicle path travel times.

The one sample K-S test results show the probe vehicle travel times follows a normal distribution while the cellular vehicle travel times do not (Table 5-14). All three nonparametric tests show that these two groups of data are from the different population.

Table 5-13
Length discrepancy and error analysis on I-895

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	10.56	0.24	0.32	3.17
Path Travel Time (sec)	787.50	685.95	106.75	12.63
Path Travel Speed (mph)	48.96	54.21	7.22	16.82

5.4.6 US 1

US Route 1 was divided into two segments: US-1a and US-1b, due to the fact that no link can be found from node 5104 to node 6195. US-1a has 23 records and US-1b has 19 records. Table 5-15 shows the error analysis on two sections of US Route 1. Both average length discrepancies are small. However, the average absolute travel time error is about 40% of the vehicular travel time on US-1a and 69.41% on US -1b. Therefore, the small length discrepancies cannot be the major reason contributing to the large difference in two types of path travel times. On both segments, the cellular data underestimate the vehicular travel times. The histogram analysis shows 20 of 23 records have absolute errors greater than 20% on US1a. On US-1b, all the records consistently have absolute errors around 70%. Table 5-16 shows the statistical tests results for US-1a and Table 5-17 shows the results for US-1b. The test results on both segments show significant difference between distributions of vehicular and cellular data.

Table 5-14
Statistical test results on I-895

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		18	18
Normal Parameters(a,b)	Mean	787.5000	687.9460
	Std. Deviation	163.43293	132.76722
Most Extreme Differences	Absolute	.208	.319
	Positive	.188	.159
	Negative	-.208	-.319
Kolmogorov-Smirnov Z		.906	1.391
Asymp. Sig. (2-tailed)		.385	.042
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-3.501	
Asymp. Sig. (2-tailed)		.000	
N(a)		19	
Chi-Square		11.842	
Df		1	
Asymp. Sig.		.001	
N(b)		19	
Kendall's W(b)		.623	
Chi-Square		11.842	
Df		1	
Asymp. Sig.		.001	

a Friedman Test
b Kendall's Coefficient of Concordance

Table 5-15
Length discrepancy and error analysis on US-1

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (a) (mile)	6.05	6.03	0.03	0.71
Path Length (b) (mile)	4.43	4.28	0.15	3.33
Path Travel Time (a) (sec)	885.78	501.68	384.10	39.99
Path Travel Time (b) (sec)	1079.95	323.43	756.52	69.41

Table 5-16
Statistical test results on US-1 a

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		23	23
Normal Parameters(a,b)	Mean	885.7826	501.6802
	Std. Deviation	367.06885	215.60040
Most Extreme Differences	Absolute	.311	.237
	Positive	.195	.119
	Negative	-.311	-.237
Kolmogorov-Smirnov Z		1.490	1.138
Asymp. Sig. (2-tailed)		.024	.150
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-4.197	
Asymp. Sig. (2-tailed)		.000	
N (a)		23	
Chi-Square		23.000	
Df		1	
Asymp. Sig.		.000	
N(b)		23	
Kendall's W(b)		1.000	
Chi-Square		23.000	
Df		1	
Asymp. Sig.		.000	

5.4.7 MD-40

41 survey runs were collected on MD-40. Table 5-18 shows the error analysis for path length and path travel times. While the path length discrepancies are small (2.53%), the travel time error is 33.54%. 33 of 41 records have absolute errors more than 50% of the vehicular travel time. Statistical tests results (Table 5-19) show significant difference between distributions of vehicular and cellular data.

5.4.8 MD-45

34 survey runs were collected on MD-45. Similar to MD-40, US-1 and other arterials, Table 5-20 shows that there are small path length discrepancies (2.53%) but large error in

travel times (58.56%). 5 of 34 records have absolute errors between 20% to 50% and the other 29 records have absolute errors more than 50% of the vehicular travel times. All three non-parametric statistical tests show significant difference between these two data sets (p=0.05), as shown in Table 5-21.

Table 5-17
Statistical test results on US-1 b

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		19	19
Normal Parameters(a,b)	Mean	1079.9474	323.4294
	Std. Deviation	280.28249	70.23323
Most Extreme Differences	Absolute	.202	.304
	Positive	.141	.182
	Negative	-.202	-.304
Kolmogorov-Smirnov Z		.882	1.324
Asymp. Sig. (2-tailed)		.417	.060
Paired Sample T Test		Vehicle-Cellular	
Z		.805	
Asymp. Sig. (2-tailed)		.000	
N (a)		19	
Chi-Square		19.000	
Df		1	
Asymp. Sig.		.000	
N (b)		19	
Kendall's W(b)		1.000	
Chi-Square		19.000	
Df		1	
Asymp. Sig.		.000	

a Friedman Test
b Kendall's Coefficient of Concordance

Table 5-18
Length discrepancy and error analysis on MD-40

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	7.00	6.87	0.18	2.53
Path Travel Time (sec)	974.05	669.45	325.56	33.54

5.4.9 Martin Luther King Boulevard

On Martin Luther King Boulevard, each survey run was around 1 mile. 38 records were collected. Table 5-22 shows the error analysis for path length and path travel times. The average absolute error is less than 0.1 mile, which is 7.8% of the actual vehicular travel distance. The average error of travel time is 47.55% of actual probe vehicle's travel time. The histogram analysis of error shows there are 9 out of 38 records in the range of [0%, 10%], 2 records have errors in the range of [10%,20%], and all the other records have errors more than 50% of the vehicular travel time. Both travel times follow the normal distribution, while the statistical tests results in Table 5-23 indicate there is significant difference between these two data sets.

In these ten groups of comparisons, statistical tests result confirmed the similarity of data on I-70 and I-695. On all the other routes, the results indicate that the distributions of two groups of travel times are not similar to each other. Though the deviation of these travel times is small, on most routes, the cellular travel times consistently underestimated the vehicular travel time. Therefore, statistically, the distributions of these two types of data are not similar.

5.5 Findings

In this chapter, we evaluated the data quality with respect to paths. Travel time was analyzed since the length discrepancies are not a major factor to cause travel time deviation. The vehicular data is deemed as "Ground Truth" in the calculation of evaluation measures.

The analyses of path travel times show that the cellular data can provide very good estimation on freeways with an average absolute error around 10%, which is smaller than the recommended error range of 20% for traveler information systems. It shows the travel time information on freeway can be readily used for potential ITS applications. On arterials, this measure increases to 40 to 60%. The percentile analysis of absolute error indicates similar pattern. Statistical tests confirm the similarity of data distributions on freeways and indicate significant difference on all arterials.

In the next chapter, we compare the cellular data, vehicular data and sensor data at 10 selected sensor locations. This comparison can be used as a reference for cellular data quality.

Table 5-19
Statistical test results on MD-40

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		41	41
Normal Parameters(a,b)	Mean	974.0488	669.4493
	Std. Deviation	241.62160	224.88774
Most Extreme Differences	Absolute	.198	.140
	Positive	.124	.082
	Negative	-.198	-.140
Kolmogorov-Smirnov Z		1.265	.899
Asymp. Sig. (2-tailed)		.081	.394
Paired Sample T Test		Vehicle-Cellular	
Z		.986	
Asymp. Sig. (2-tailed)		.000	
N (a)		41	
Chi-Square		33.390	
Df		1	
Asymp. Sig.		.000	
N (b)		41	
Kendall's W(b)		.814	
Chi-Square		33.390	
Df		1	
Asymp. Sig.		.000	

a Friedman Test
b Kendall's Coefficient of Concordance

Table 5-20
Length discrepancy and error analysis on MD-45

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	4.58	4.53	0.09	2.16
Path Travel Time (sec)	955.52	381.10	574.42	58.56

Table 5-21
Statistical test results on MD 45

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		34	34
Normal Parameters(a,b)	Mean	931.7941	371.9726
	Std. Deviation	351.45203	129.48716
Most Extreme Differences	Absolute	.139	.255
	Positive	.105	.136
	Negative	-.139	-.255
Kolmogorov-Smirnov Z		.810	1.488
Asymp. Sig. (2-tailed)		.528	.024
Wilcoxon Signed Ranks Test		Vehicle-Cellular	
Z		-5.086	
Asymp. Sig. (2-tailed)		.000	
N (a)		34	
Chi-Square		34.000	
Df		1	
Asymp. Sig.		.000	
N (b)		34	
Kendall's W(b)		1.000	
Chi-Square		34.000	
Df		1	
Asymp. Sig.		.000	

a Friedman Test
b Kendall's Coefficient of Concordance

Table 5-22
Length discrepancy and error analysis on Martin Luther King Boulevard

	Avg. Vehicular	Avg. Cellular	Avg. Abs. Error	Avg. % of Absolute Error
Path Length (mile)	1.20	1.21	0.09	7.80
Path Travel Time (sec)	239.11	186.78	109.81	47.55

Table 5-23
Statistical test results on Martin Luther King Boulevard

One-Sample Kolmogorov-Smirnov Test		Probe Vehicle	Cellular Phone
N		38	38
Normal Parameters(a,b)	Mean	239.1053	186.7758
	Std. Deviation	50.25975	129.45671
Most Extreme Differences	Absolute	.179	.183
	Positive	.179	.183
	Negative	-.094	-.164
Kolmogorov-Smirnov Z		1.101	1.129
Asymp. Sig. (2-tailed)		.177	.156
Paired Sample T Test		Vehicle-Cellular	
Z		0.015	
Asymp. Sig. (2-tailed)		.930	
N (a)		38	
Chi-Square		8.526	
Df		1	
Asymp. Sig.		.004	
N (b)		38	
Kendall's W(b)		.224	
Chi-Square		8.526	
Df		1	
Asymp. Sig.		.004	

a Friedman Test

b Kendall's Coefficient of Concordance

6 Comparison with Fixed Sensor Data

6.1 Data Preparation

As described in Chapter 3, traffic speed data in 5-minute intervals are downloaded at 10 sensor locations, from 26 January to 3 February 2006. The speed data are downloaded in the format of Microsoft EXCEL table file from the following website: http://www.cattlab.umd.edu/cf/index.cfm?js=enabled&bin=trafficData

The sensor traffic speeds are spot speeds. They represent the average traffic speeds at the locations of sensor at a given time. The cellular speeds represent the average traffic speed on a link in a given 5-minute time interval. To compare with the sensor speed data, the corresponding cellular traffic speeds during the same time interval (26 January to 3 February 2006) were queried from the cellular probe data set. The following steps are performed to get the cellular data prepared for comparison:
Step 0: match location of the sensor to the TMC node number, determine traffic direction;
Step 1: run query in cellular database to retrieve the cellular speed data in the same time interval and direction on links whose end node or start node is the sensor location's corresponding TMC node number.
Step 2: calculate weighted average link travel speeds from cellular link speeds.
Step 3: find the 5-minute time interval of cellular traffic speed which contains the timestamp of the sensor traffic speed.

For example, the sensor on Interstate 695 inner loop at Stevenson Road can be matched to the TMC node 4526 in direction 1 (the location of the sensor is shown in Figure 6-1).

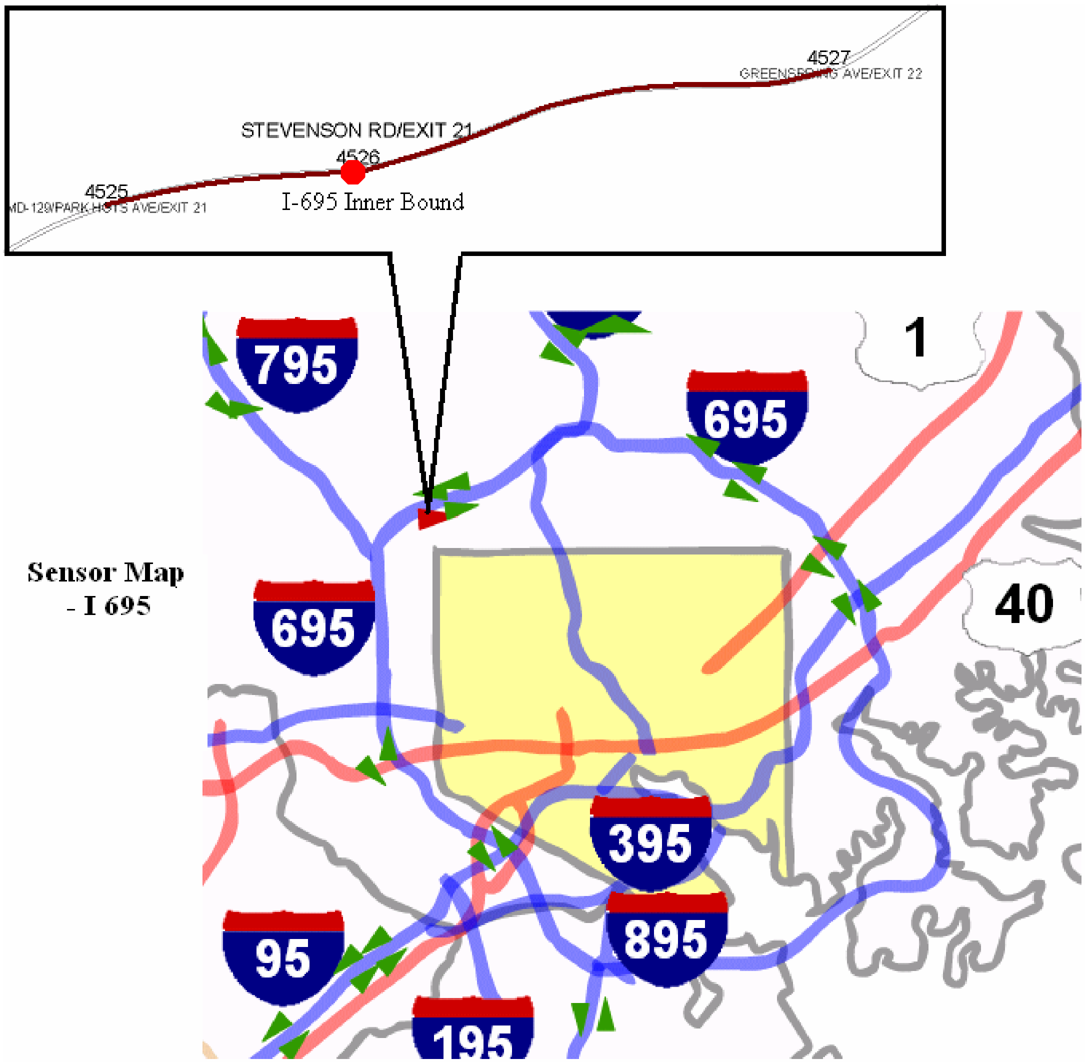

Figure 6-1
Location of sensor: I-695 outer loop @ Stevenson Road

To compare with the sensor speed (71 mile/hour) at time 0:35 AM on 26 January, 2006, two cellular speeds are retrieved. The cellular speeds on the TMC link 4526 in direction 1 at time 0:35 AM on 26 January 2006 is 103.13 kilometer per hour and the speed on the TMC link 4527 in direction 1 at time 0:35 AM on 26 January 2006 is 103.74 kilometer per hour . The length of link 4526 in direction 1 is 1443 meter and the length of link 4527 in direction 1 is 2267 meter. The weighted cellular speed will be calculated as shown below:

$$
\begin{aligned}
Weighted\_Cellular\_Speed &= (103.13*1443+103.74*2267)/(1443+2267) \\
&= 103.50\ kilometer/hour \qquad (6\text{-}1) \\
&= 64.31\ mile/hour
\end{aligned}
$$

Since the lengths of these two adjacent links are short, we use the weighted average link traffic speed to approximate the spot traffic speed at node 4526.

When the vehicular probe data is available in the same time interval of interest, the same calculation can be done to derive the weighted vehicular speed. When the travel time of probe vehicles traverse more than one 5-minute interval on a link. An interpolation procedure such as the one described in Chapter 4 will be applied to obtain corresponding cellular speeds and sensor speeds.

6.2 Travel Speed Comparison

As discussed before, due to the limited availability of sensor data, 10 sensor locations were selected for this comparison. Since sensor data only provides spot speeds, it is not feasible to compare travel times on links or paths, but travel speeds at selected locations. In the following sections, the results of comparison are presented based on the sensor locations.

6.2.1 I-695 Outer Loop @ US 40 West

This sensor location is matched to the TMC node 4519 in direction 0. The cellular data is available from 0:00:00 on 26 January to 24:00:00 on 3 February 2006, while the sensor data is available from 0:00:00 on 26 January to 00:09:00 on 31 January 2006. The variations of cellular and sensor travel speeds in response to time of day are illustrated below for 27 January 2006, when 25 probe vehicle speeds are available spanning from 8:10 AM to 6:32 PM. Figure 6-2 shows the value of speeds from all three data sources on the 27th, while Figure 6-3 focuses on the time interval during a day in which all three types of data are available.

Based on the probe vehicle speeds, Figure 6-4 shows the deviations of cellular-vehicular speeds and sensor-vehicular ones. It is observed that, at this location, the sensor speed is faster than the probe vehicle speed in 23 out of 25 cases. For cellular data, in 12 out of 25 cases, the cellular probe speed is larger than the probe vehicle speed, leading to a smaller average deviation for the cellular data. However, the absolute deviation between sensor speeds and vehicular ones is smaller than that between cellular speeds and vehicular ones. Table 6-1 summarizes the speed deviations.

The following steps are taken to evaluate the statistical correlation between the three data sources. First, a non-parametric One-Sample Kolmogorov-Smirnov test is selected to test whether all three data are normally distributed. Depending on this test results, either a parametric or a non-parametric standard statistical test is utilized to determine the correlation between cellular, sensor, and probe vehicle speeds. The paired comparisons are denoted as cellular-vehicular, sensor-vehicular and cellular-sensor. Finally test results are returned for interpretation.

For this sensor location, cellular and sensor data demonstrates a possible normal distribution in the first step test. Therefore, a t-test (parametric) is applicable to test the association between cellular and sensor data. The results demonstrate similar distribution between them. A non-parametric Wilcoxon Signed Rank Test is selected to test the correlation between the other pair of data sources. It is observed that a strong correlation exists between cellular and probe vehicle speeds, while sensor speeds and vehicular data show significant difference in distribution. Table 6-2 shows the statistics of travel speed comparison at the TMC node 4519 in direction 0, corresponding to location of I-695 at US 40 west.

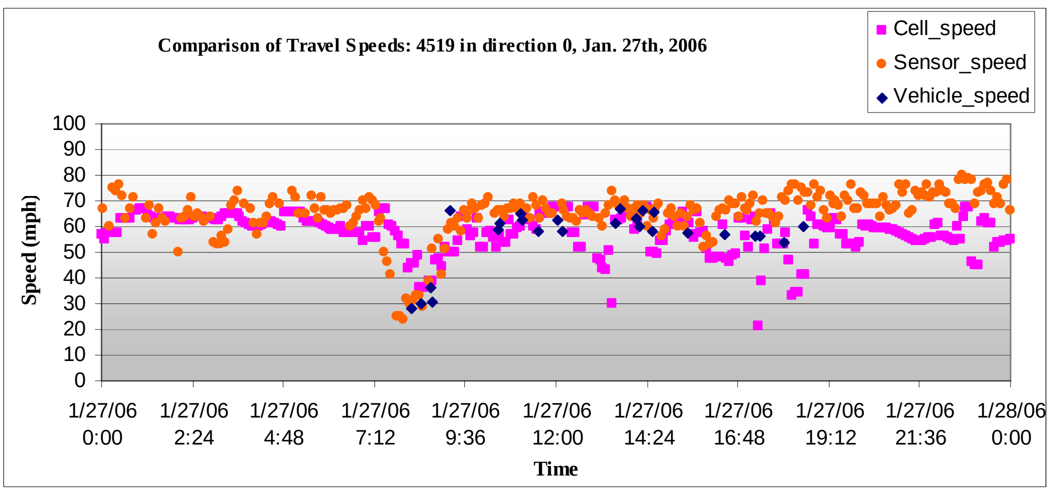

Figure 6-2
Comparison of travel speeds: I-695 outer loop @ US 40 West, 27 January 2006

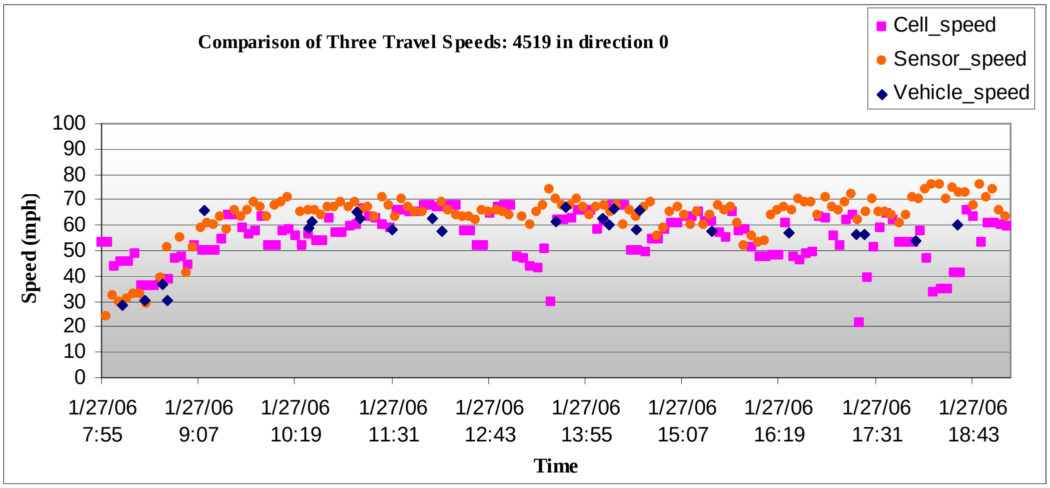

Figure 6-3
Comparison of three travel speeds: I-695 outer loop @ US 40 West, 27 January 2006

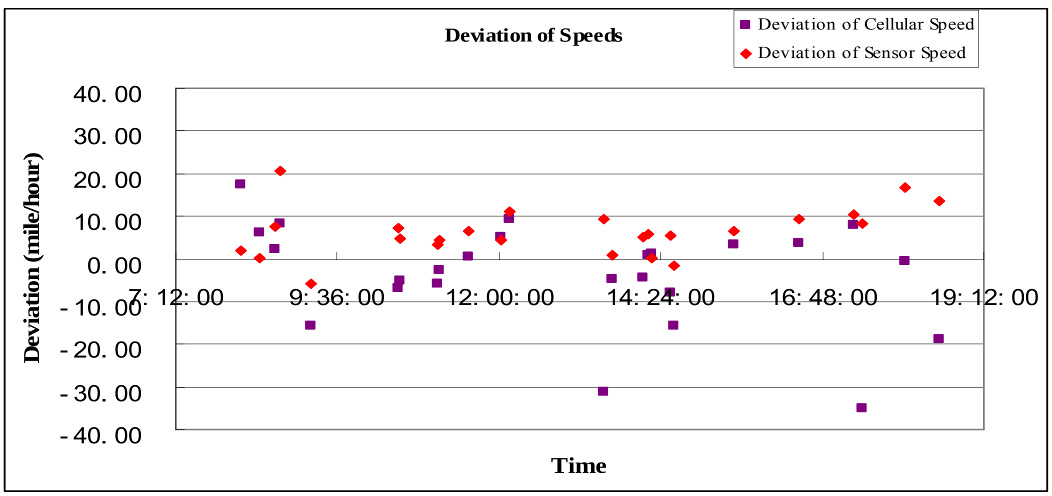

Figure 6-4
Speed deviations: I-695 outer loop @ US 40 West, 27 January 2006

Table 6-1
Speed deviations: I-695 outer loop @ US 40 West

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-3.55	8.85	6.22	6.80

6.2.2 I-695 @ I-70 Inner Loop

In this example, the sensor location is matched to the TMC node 4520 in direction 1. The cellular data is available from 0:00:00 on 26 January to 23:55:00 on 3 February, 2006, while the sensor data is available from 0:00:00 on 26 January to 00:07:00 on 31 January 2006. On 27 January 26 probe vehicle speeds are available spanning from 7:12 AM to 5:47 PM. Figure 6-5 shows the speed variation from the three data sources for the day, while Figure 6-6 focuses on the time period when all three types of data are available. Based on the probe vehicle speeds, Figure 6-7 shows the deviations of cellular-vehicular speeds and sensor-vehicular ones.

For this location, sensor speeds are smaller than the ones from the probe vehicle in 11 out of 26 cases. It is shown that the sensor speed spreads more evenly as compared to the previous example, resulting in a significant reduction in the averaged deviation (from 6.22 to 1.69). However, cellular data still demonstrates a larger absolute deviation from the sensor data. On average, cellular speeds are larger than the probe vehicle ones at this location. Table 6-3 summarizes the speed deviation at the TMC node 4520 in direction 1.

Table 6-2
Statistics of travel speed comparison: I-695 outer loop @ US 40 West

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		25	25	25
Normal Parameters(a,b)	Mean	56.0440	52.4980	62.2620
	Std. Deviation	11.59300	12.24948	11.31832
Most Extreme Differences	Absolute	.299	.169	.341
	Positive	.171	.109	.191
	Negative	-.299	-.169	-.341
Kolmogorov-Smirnov Z		1.494	.845	1.704
Asymp. Sig. (2-tailed)		.023	.473	.006
Paired Sample Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		-.928(b)	.389(a)	-3.942(b)
Asymp. Sig. (2-tailed)		.353	.055	.000

Paired sample t-Test
Wilcoxon Signed Ranks Test

One-Sample Kolmogorov-Smirnov test results show a less possibility for cellular data to follow normal distribution. Statistical test shows that all three data sources have similar distribution to each other, where cellular-vehicle pair demonstrates the closest correlation. Table 6-4 provides the statistical summary. At this location, sensor data outperformed the cellular probe data in the sense of average deviation, average absolute deviation and similarity of data distribution.

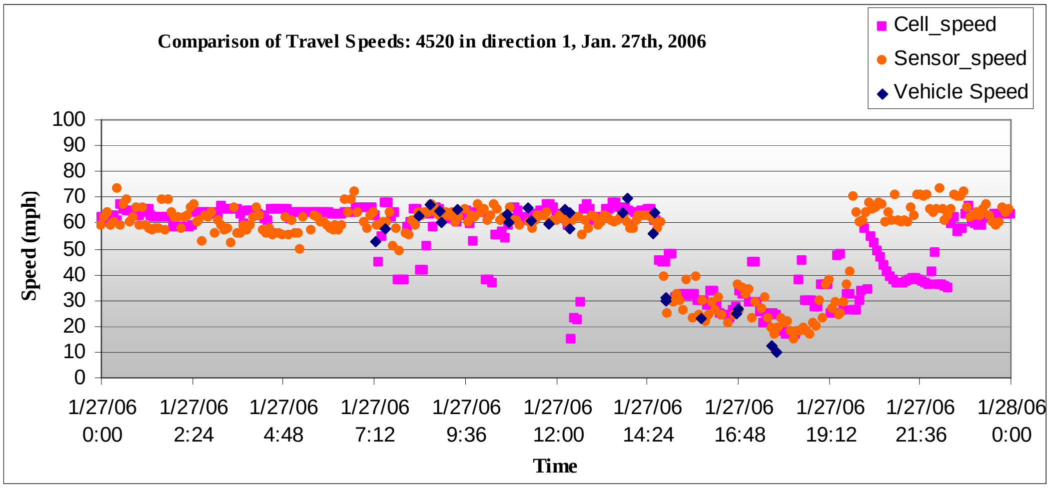

Figure 6-5
Comparison of travel speeds: I-695 inner loop @ I-70, 27 January 2006

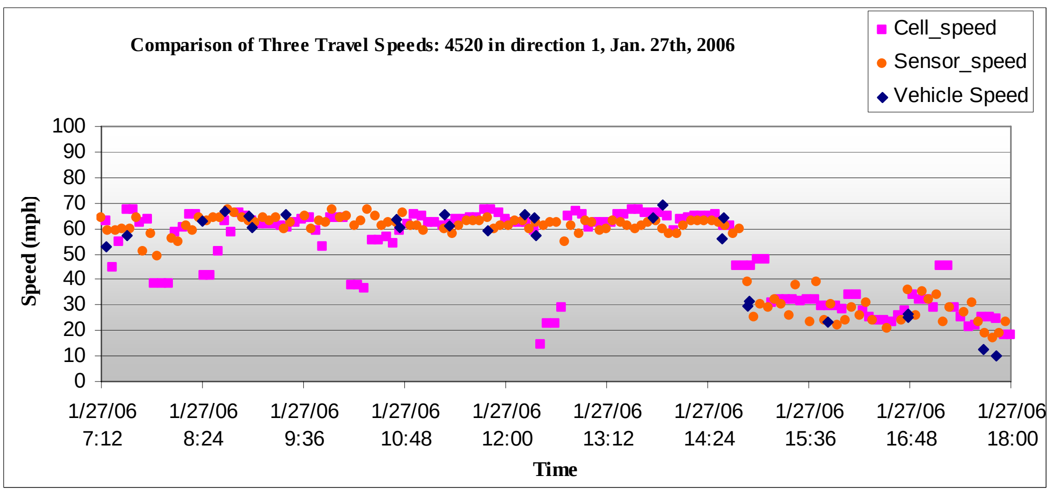

Figure 6-6
Comparison of three travel speeds: I-695 inner loop @ I-70, 27 January 2006

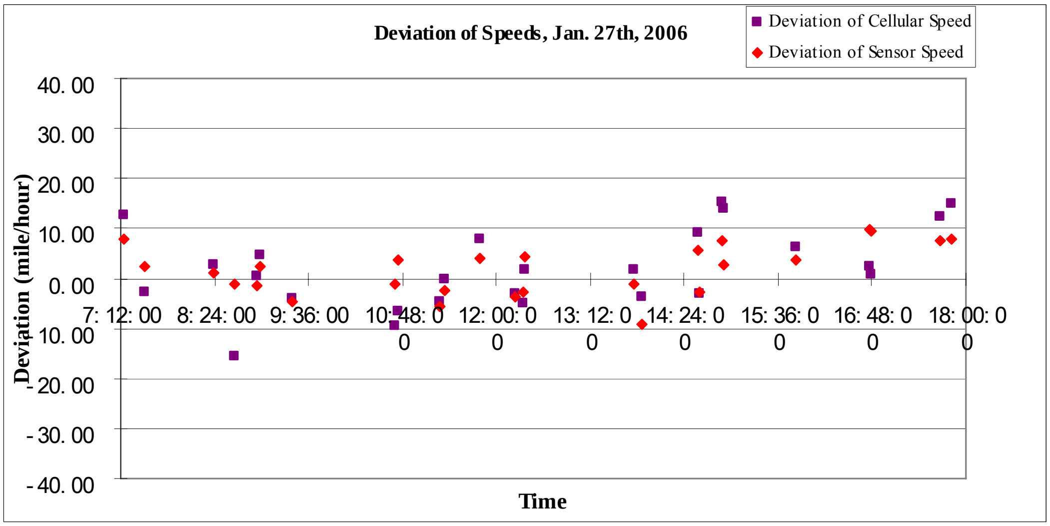

Figure 6-7
Speed deviations: I-695 inner loop @ I-70, 27 January 2006

Table 6-3
Speed deviations: I-695 inner loop @ I-70

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
1.90	6.36	1.69	4.46

6.2.3 I-695 Inner Loop @ Stevenson Road

This sensor location is matched to the TMC node 4526 in direction 1. It corresponds to the location of I-695 Inner Loop at Stevenson Road. The cellular data is available from 0:00:00 on 26 January to 23:55:00 on 3 February 2006, while the sensor data is available from 0:00:00 on 26 January to 23:55:00 on 3 February 2006. Available sensor data on 3 February enables us to compare all three data sources for an additional day. On 27 January 28 probe vehicle speeds are available and another 12 speeds are available on 3 February.

Table 6-4
Statistics of travel speed comparison: I-695 inner loop @ I-70

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		26	26	26
Normal Parameters(a,b)	Mean	51.4931	53.3923	53.1858
	Std. Deviation	18.63554	14.46719	15.18727
Most Extreme Differences	Absolute	.286	.230	.375
	Positive	.169	.167	.206
	Negative	-.286	-.230	-.375
Kolmogorov-Smirnov Z		1.460	1.172	1.913
Asymp. Sig. (2-tailed)		.028	.128	.001
Wilcoxon Signed Ranks Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		-.825	-.622	-1.613
Asymp. Sig. (2-tailed)		.409	.534	.107

Figure 6-8 shows the speed variations on 27 January, while Figure 6-9 presents the ones for 3 February. Figures 6-10 and 6-11 demonstrate the data from all three data sources with more details in corresponding time intervals. Figures 6-12 and 6-13 show the deviations of cellular-vehicular speeds and sensor-vehicular speeds for both available time intervals. The data on 3 February suggests that the sensor probe speeds are always larger than the vehicular probe speeds while the cellular probe speeds are always smaller than the vehicular ones.

At this location, both cellular and sensor deviations are evenly spread across the x-axis. On average, cellular data underestimates the vehicle speed with a larger absolute deviation, while sensor data shows a larger average deviation and a smaller absolute one. Table 6-5 shows the speed deviation for I-695 Inner Loop at Stevenson Rd.

Statistically, all three data are normally distributed. Paired t-tests are used to examine the differences between the data. It is observed that all variables have a distribution that is significantly different from the others. Table 6-6 summarizes the statistics of the travel speed comparison at the TMC node 4526, direction 1.

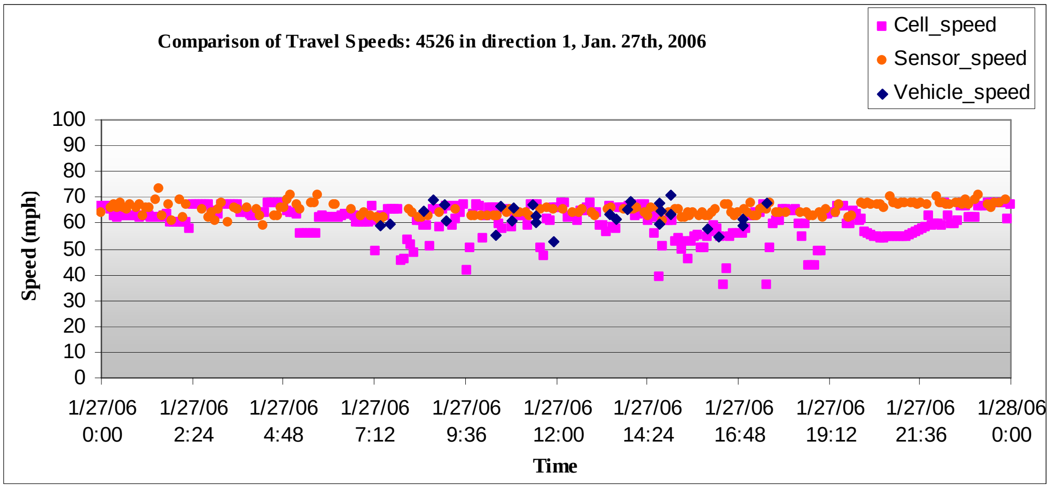

Figure 6-8
Comparison of travel speeds: I-695 inner loop @ Stevenson Rd., 27 January 2006

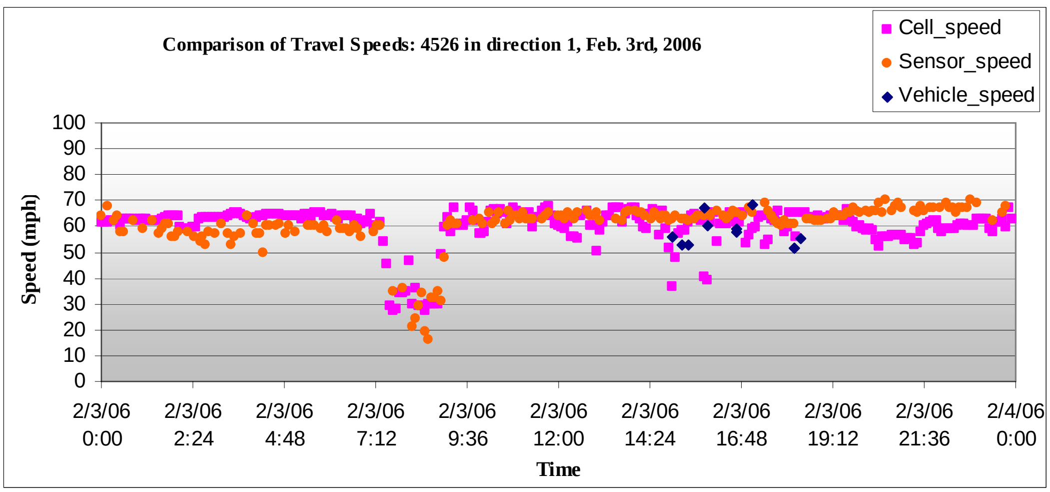

Figure 6-9
Comparison of travel speeds: I-695 Inner Loop @ Stevenson Rd., 3 February 2006

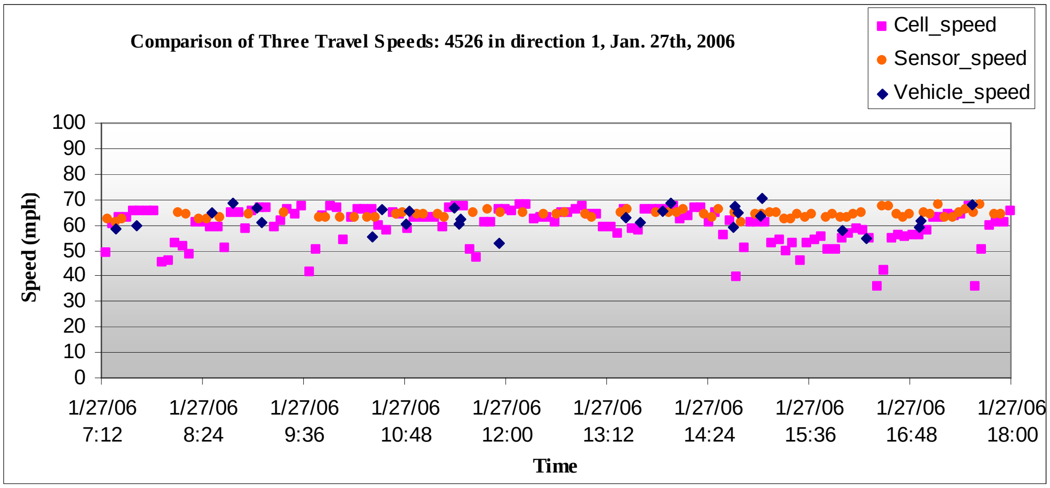

Figure 6-10
Comparison of Three Travel Speeds: I-695 Inner Loop @ Stevenson Rd., 27 January 2006

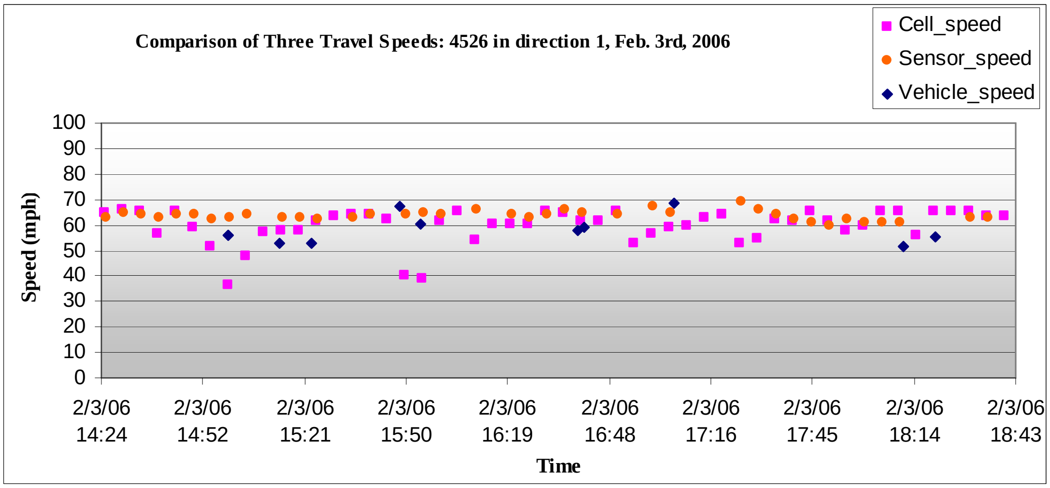

Figure 6-11
Comparison of Three Travel Speeds: I-695 Inner Loop @ Stevenson Rd., 3 February 2006

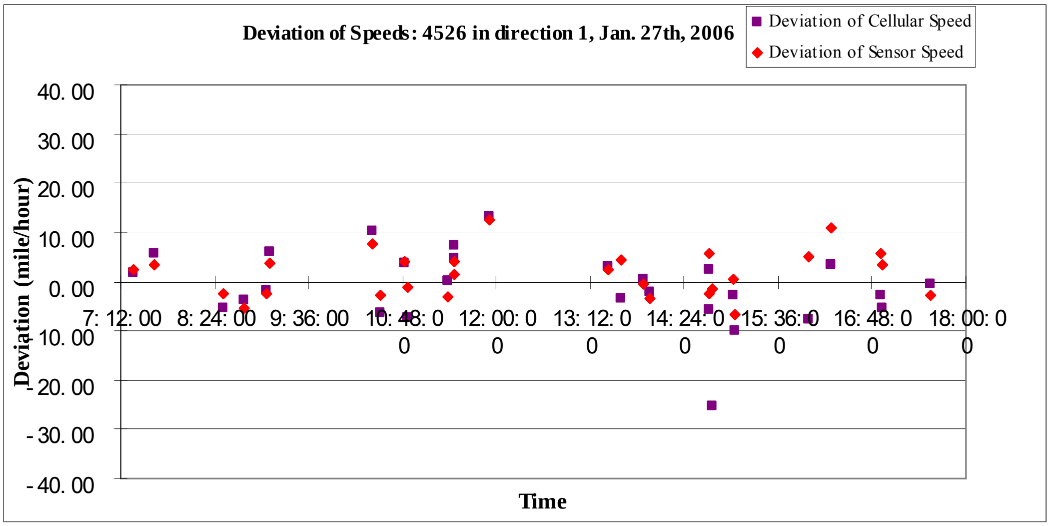

Figure 6-12
Speed Deviations: I-695 Inner Loop @ Stevenson Rd., 27 January 2006

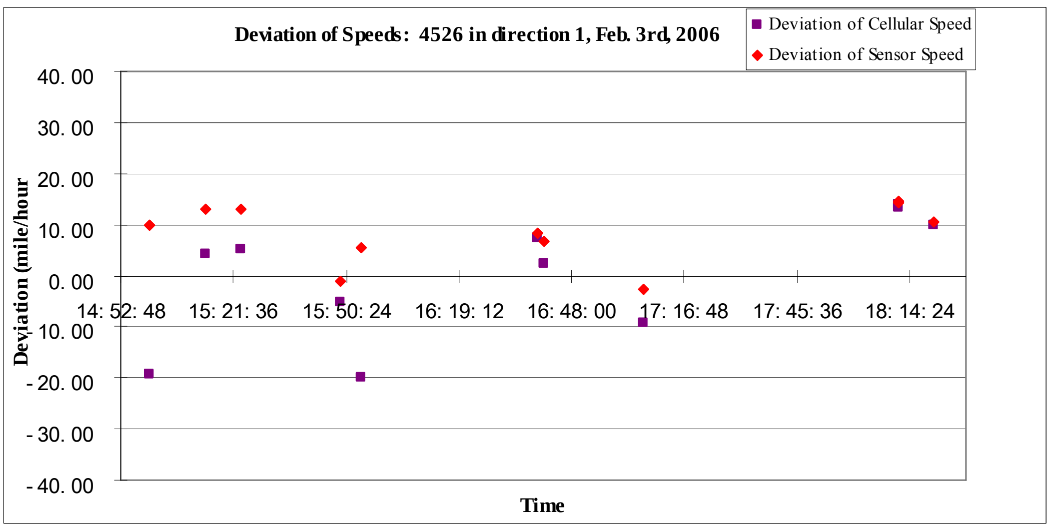

Figure 6-13
Speed Deviations: I-695 Inner Loop @ Stevenson Rd., 3 February 2006

Table 6-5
Speed deviations: I-695 inner loop @ Stevenson Rd.

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-1.10	7.06	3.62	5.54

Table 6-6
Statistics of travel speed comparison: I-695 inner loop @ Stevenson Rd.

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		40	40	40
Normal Parameters(a,b)	Mean	59.9842	64.7093	53.1858
	Std. Deviation	8.33793	1.24628	15.18727
Most Extreme Differences	Absolute	.204	.150	.375
	Positive	.187	.150	.206
	Negative	-.204	-.150	-.375
Kolmogorov-Smirnov Z		.595	1.287	.950
Asymp. Sig. (2-tailed)		.870	.073	.328
Paired Sample T Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		.475	.389	.876
Asymp. Sig. (2-tailed)		.016	.055	.000

6.2.4 I-695 Outer Loop @ Stevenson Rd.

This location is the same as the previous one on outer loop, therefore the sensor location is matched to the TMC node 4526 in direction 0. The cellular data is available for the whole day from 26 January to 3 February 2006. Probe vehicle data is available for 28 points on 27 January and 8 points are available for 3 February. Note that sensor data is only available from 11:42:00 on 27 January to 23:50:00 on 3 February 2006. This only allows us to match 24 sensor speeds to the vehicle data.

The speed variations for all three data sources with respect to the time of day are illustrated in Figures 6-14 and 6-15, for 27 January and 3 February respectively. Figures 6-16 and 6-17 focus on the time interval during a day in which all three types of data are available. Figures 6-18 and 6-19 show the deviations of cellular-vehicular speeds and sensor-vehicular ones. Unavailable sensor data prior to 11:42:00 on 27 January is clearly observable from Figures 6-14 and 6-16.

Table 6-7 summarizes the speed deviation at the TMC node 4526 in direction 0. Both data overestimate the vehicular speed and sensor data demonstrates a larger absolute deviation. The cellular probe data outperform the sensor data in the sense of average deviation and average absolute deviation. Statistical test results are presented in Table 6-8. It is observed that all three data are likely normally distributed. Paired t-test results show that the difference between cellular-vehicle and sensor-vehicle data are statistically significant.

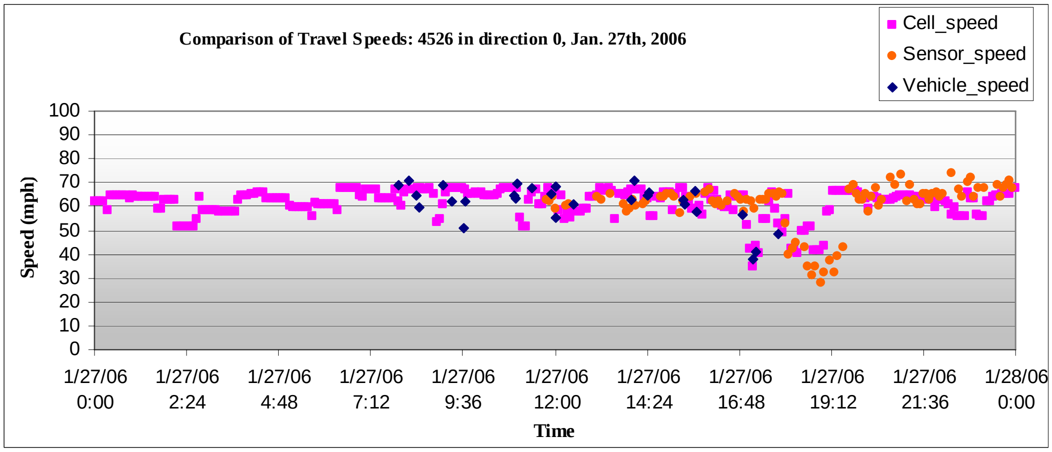

Figure 6-14
Comparison of Travel Speeds: I-695 Outer Loop @ Stevenson Rd., 27 January 2006

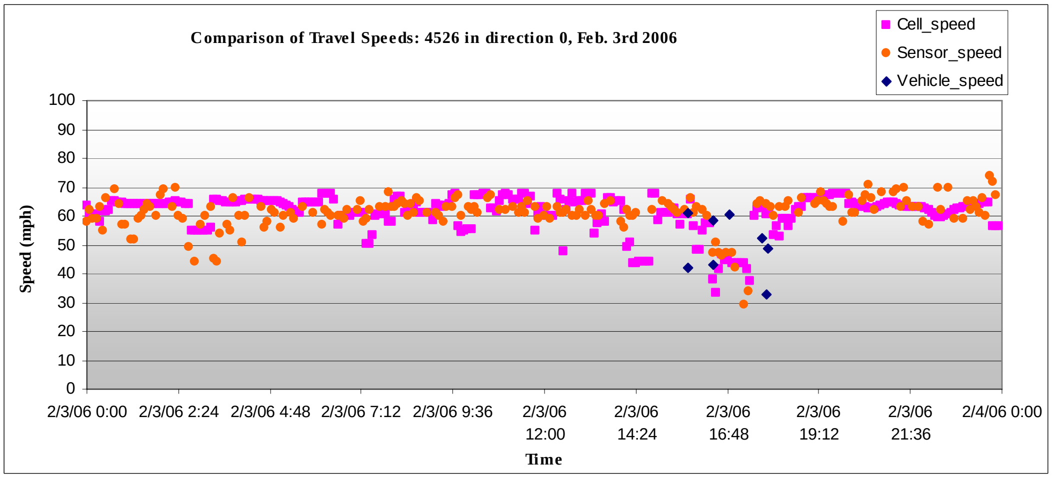

Figure 6-15
Comparison of Travel Speeds: I-695 Outer Loop @ Stevenson Rd., 3 February 2006

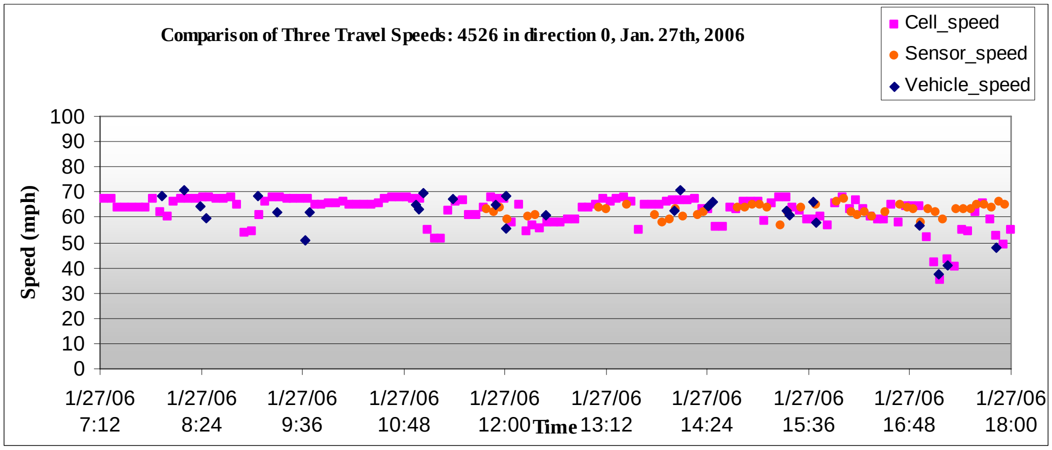

Figure 6-16
Comparison of Three Travel Speeds: I-695 Outer Loop @ Stevenson Rd., 27 January 2006

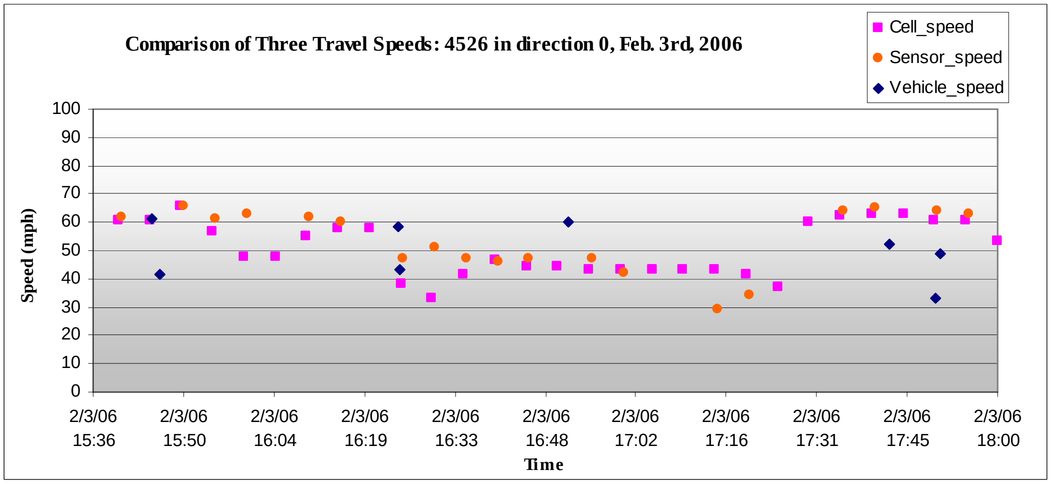

Figure 6-17
Comparison of Three Travel Speeds: I-695 Outer Loop @ Stevenson Rd., 3 February 2006

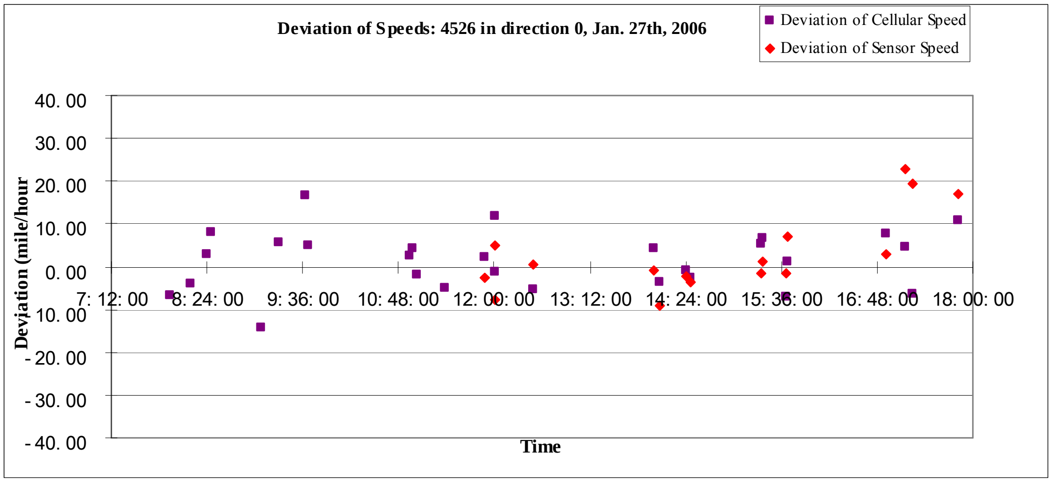

Figure 6-18
Speed Deviations: I-695 Outer Loop @ Stevenson Rd., 27 January 2006

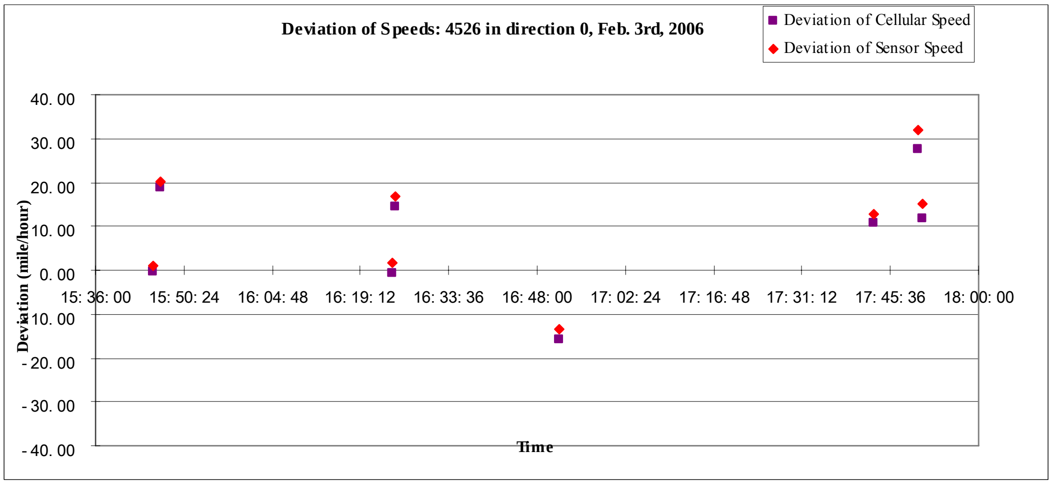

Figure 6-19
Speed Deviations: I-695 Outer Loop @ Stevenson Rd., 3 February 2006

Table 6-7
Speed deviations: I-695 outer loop @ Stevenson Rd.

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
2.88	7.10	5.34	8.91

6.2.5 I-695 Inner Loop @ Joppa Rd.

This sensor location is matched to the TMC node 4534 in direction 1, and it corresponds to the location at I-695 inner loop at Joppa Rd. The cellular data is available from 26 January to 3 February 2006, while the sensor data is available from 00:55:00 on 26 January to 23:55:00 on 3 February 2006. Twenty-eight probe vehicle data points are available on 27 January and twelve points are available on 3 February. Figures 6-20 and 6-21 demonstrate the speed variations for all three data sources with respect to the time of day, with a more detailed plot displayed in Figures 6-22 and 6-23. Based on the probe vehicle speeds, Figures 6-24 and 6-25 show the speed deviations of cellular-vehicular and sensor-vehicular.

Table 6-8
Statistics of travel speed comparison: I-695 outer loop @ Stevenson Rd.

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		24	24	24
Normal Parameters(a,b)	Mean	59.9071	61.5175	53.1858
	Std. Deviation	8.47159	3.56746	15.18727
Most Extreme Differences	Absolute	.227	.266	.375
	Positive	.176	.154	.206
	Negative	-.227	-.266	-.375
Kolmogorov-Smirnov Z		.790	1.111	1.301
Asymp. Sig. (2-tailed)		.560	.169	.068
Paired Sample T Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		.523	.391	-.122
Asymp. Sig. (2-tailed)		.009	.059	.570

Table 6-9 shows the speed deviations at the TMC node 4534 in direction 1. It shows clearly that sensor data overestimates while cellular data underestimates the vehicle speeds while the cellular data better match the trend of vehicular speed variations. Sensor data has a larger average absolute deviation for this example. Cellular probe data outperforms sensor data in this location based on deviation.

The statistical comparison of the three speeds (Table 6-10) shows that all three data conform to the normal distribution. Paired-sample t-tests show that cellular speeds have a strong correlation to the vehicle ones.

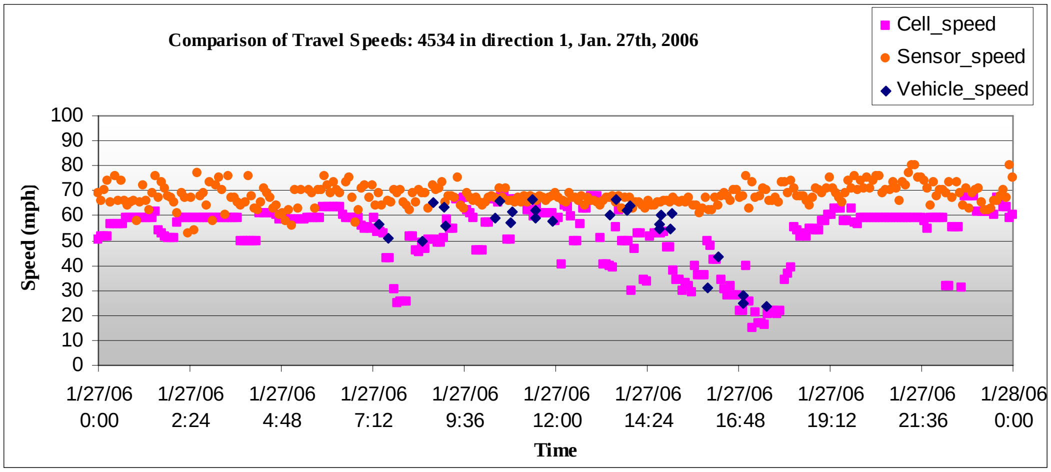

Figure 6-20
Comparison of Travel Speeds: I-695 Inner Loop @ Joppa Rd., 27 January 2006

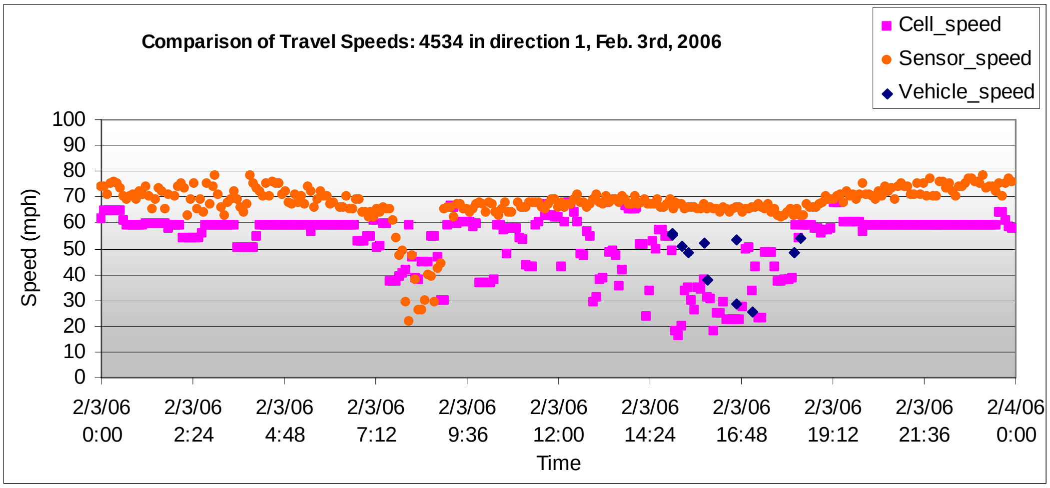

Figure 6-21
Comparison of Travel Speeds: I-695 Inner Loop @ Joppa Rd., 3 February 2006

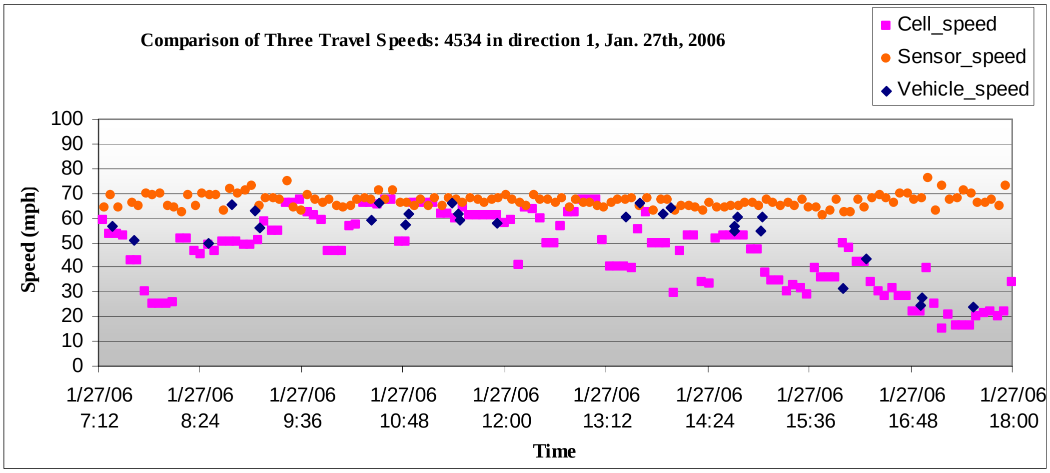

Figure 6-22
Comparison of Three Travel Speeds: I-695 Inner Loop @ Joppa Rd., 27 January 2006

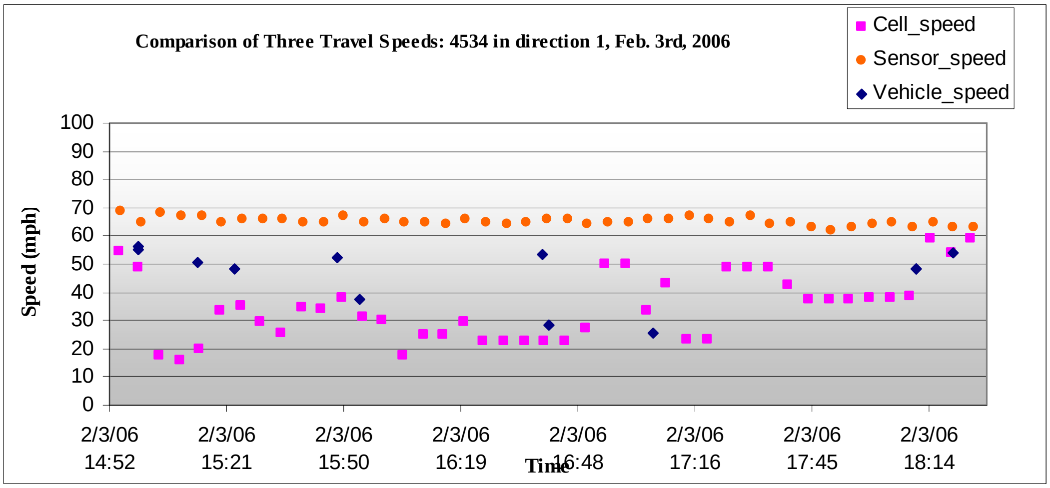

Figure 6-23
Comparison of Three Travel Speeds: I-695 Inner Loop @ Joppa Rd., 3 February 2006

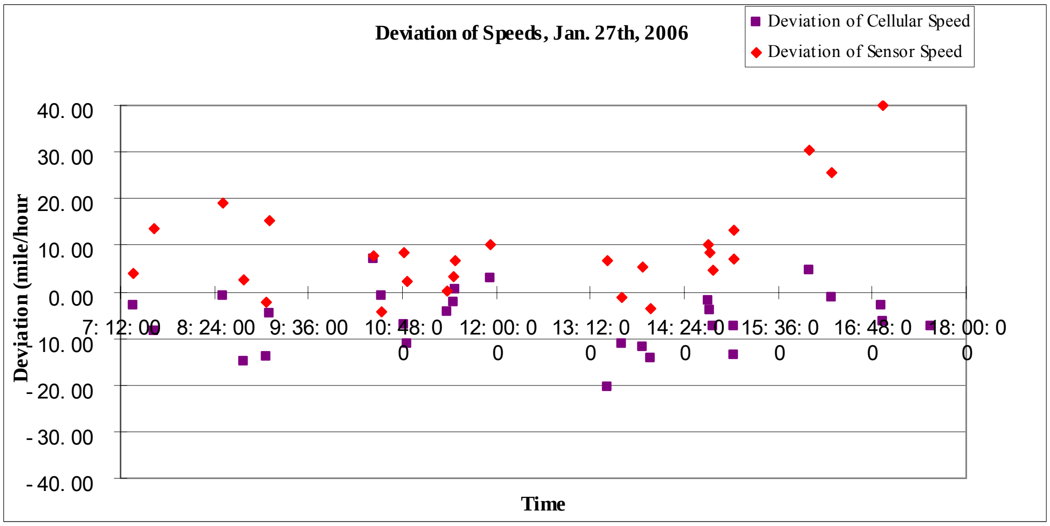

Figure 6-24
Speed Deviations: I-695 Inner Loop @ Joppa Rd., 27 January 2006

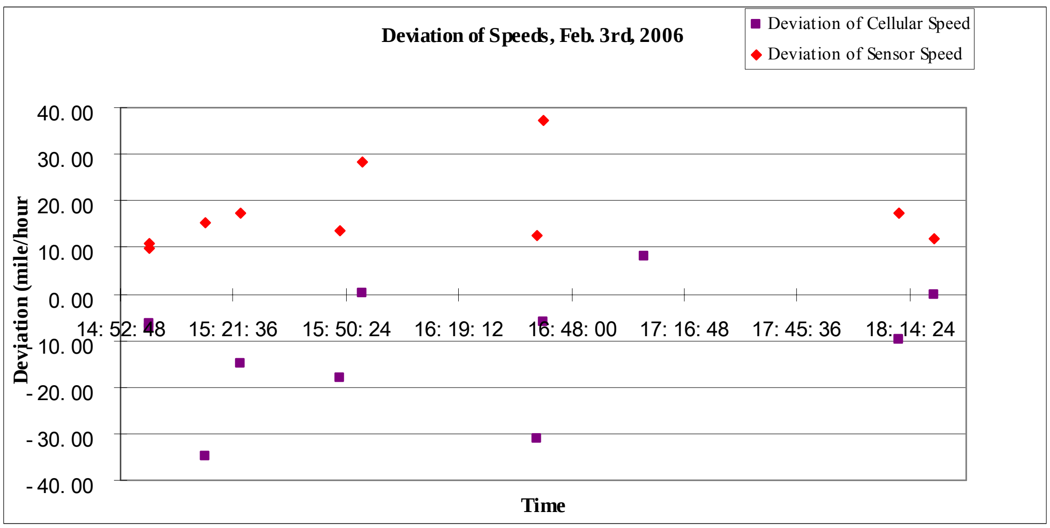

Figure 6-25
Speed Deviations: I-695 Inner Loop @ Joppa Rd., 3 February 2006

Table 6-9
Speed deviations: I-695 inner loop @ Joppa Rd.

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-7.40	8.58	13.77	14.33

Table 6-10
Statistics of travel speed comparison: I-695 inner loop @ Joppa Rd.

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		40	40	40
Normal Parameters(a,b)	Mean	44.6285	65.7903	53.1858
	Std. Deviation	13.28071	2.24675	15.18727
Most Extreme Differences	Absolute	.175	.187	.375
	Positive	.102	.138	.206
	Negative	-.175	-.187	-.375
Kolmogorov-Smirnov Z		1.184	1.106	1.184
Asymp. Sig. (2-tailed)		.121	.173	.121
Paired Sample t-Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		0.766	-.140	-.169
Asymp. Sig. (2-tailed)		.000	.389	.289

6.2.6 I-695 Outer Loop @ Joppa Rd.

This sensor location is matched to the TMC node 4535 in direction 0. The cellular data is available from 26 January to 3 February 2006, while the sensor data is available from 00:55:00 on 26 January to 23:55:00 on 3 February 2006. Twenty-eight probe vehicle data is available for 27 January and another eight points are available for 3 February. The speed variations for all three data sources with respect to the time of day are illustrated in Figures 6-26 and 6-27. Figures 6-28 and 6-29 focus on the time interval during a day in which all three types of data are available. Figures 6-30 and 6-31 show the speed deviations of cellular and sensor data compared to the vehicular speeds.

Table 6-11 summarizes the speed deviation at the TMC node 4535 in direction 0. Same trend is observed as in example 5.2.5. Sensor data overestimates the vehicular speed and cellular data underestimates it.

Table 6-12 summarizes the statistical test results on the travel speed comparisons. None of the three data are normally distributed. Wilcoxon Signed Ranks Test results suggest that the differences among these three data sources distributions are statistically significant. However, due to the smaller average deviation and average absolute deviation, the sensor data can better represents the vehicle travel speed.

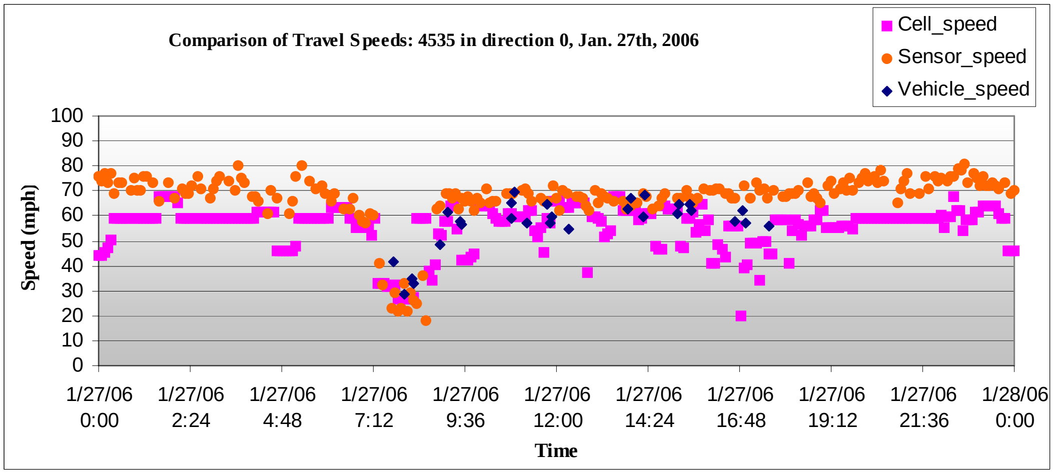

Figure 6-26
Comparison of Travel Speeds: I-695 Outer Loop @ Joppa Rd., 27 January 2006

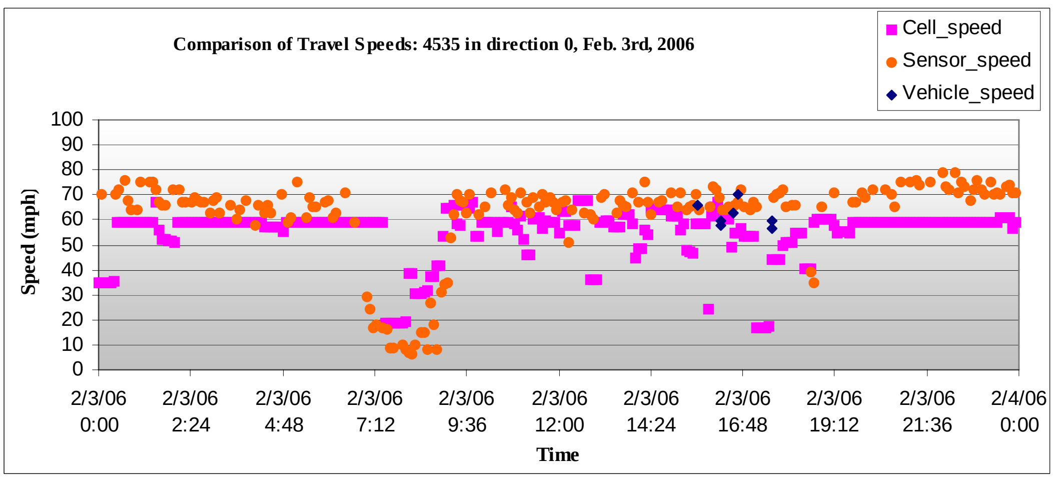

Figure 6-27
Comparison of Travel Speeds: I-695 Outer Loop @ Joppa Rd., 3 February 2006

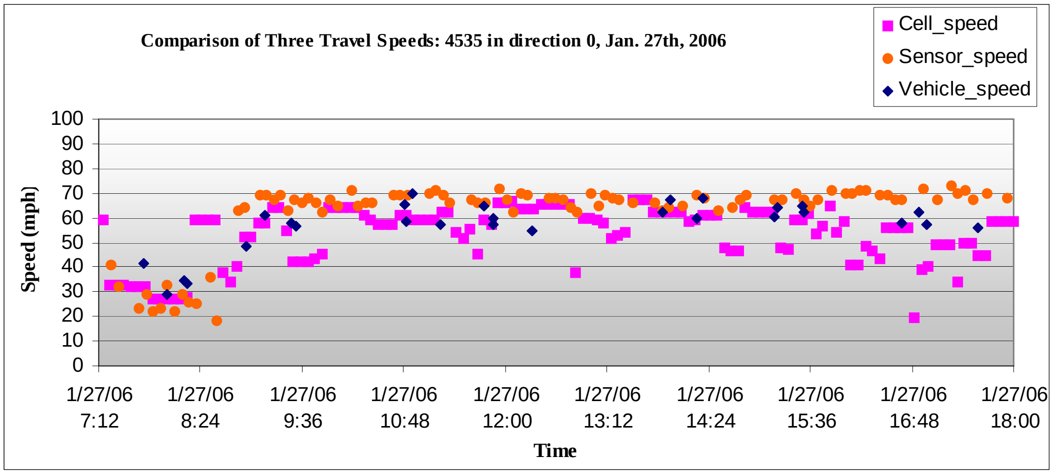

Figure 6-28
Comparison of Three Travel Speeds: I-695 Outer Loop @ Joppa Rd., 27 January 2006

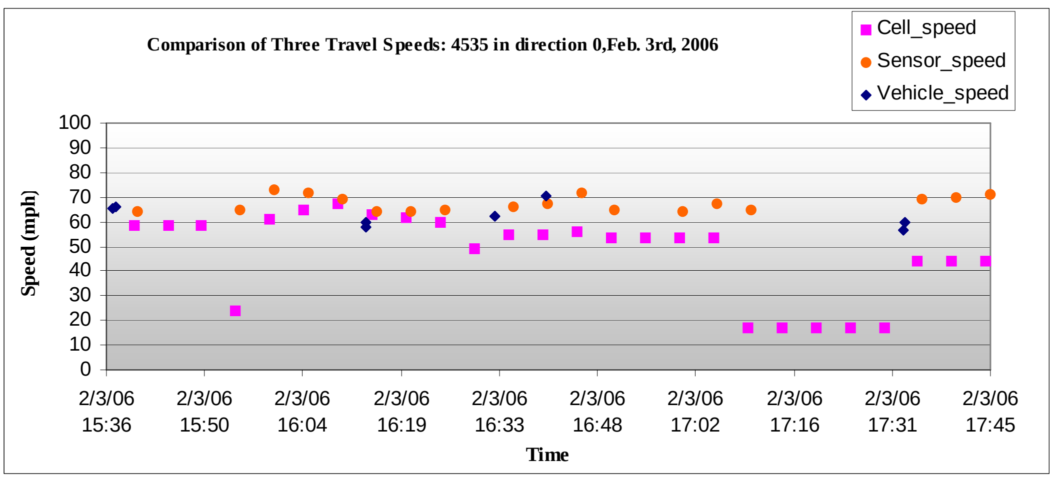

Figure 6-29
Comparison of Three Travel Speeds: I-695 Outer Loop @ Joppa Rd., 3 February 2006

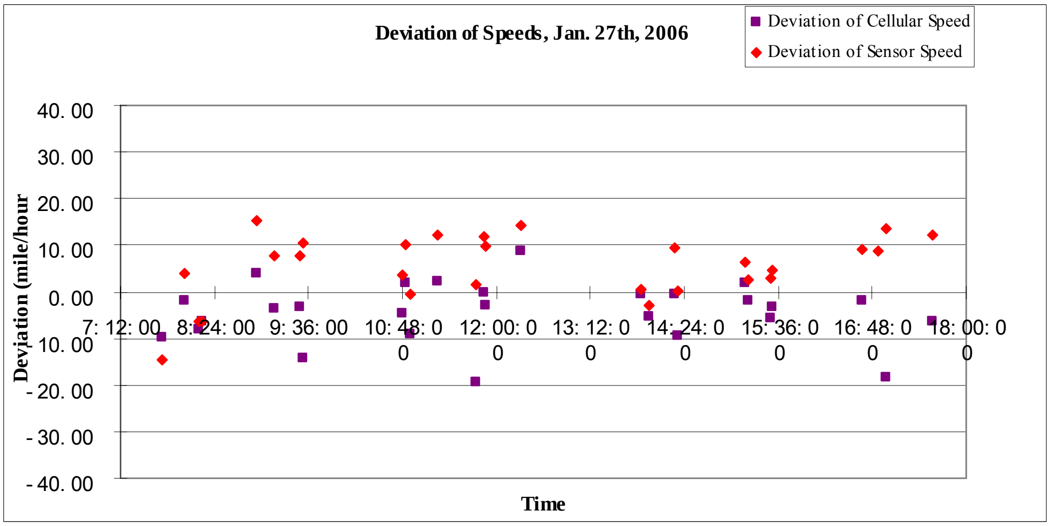

Figure 6-30
Speed Deviations: I-695 Outer Loop @ Joppa Rd., 27 January 2006

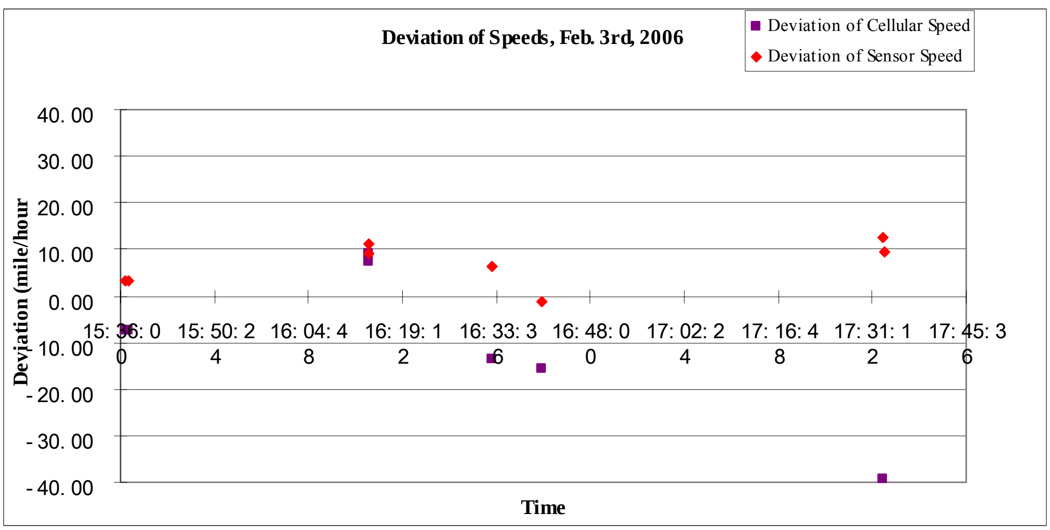

Figure 6-31
Speed Deviations: I-695 Outer Loop @ Joppa Rd., 3 February 2006

Table 6-11
Speed deviations: I-695 outer loop @ Joppa Rd.

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-7.48	9.48	5.61	7.40

Table 6-12
Statistics of travel speed comparison: I-695 inner loop @ Joppa Rd.

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		36	36	36
Normal Parameters(a,b)	Mean	50.5702	63.6603	53.1858
	Std. Deviation	14.84606	12.68612	15.18727
Most Extreme Differences	Absolute	.244	.372	.375
	Positive	.136	.281	.206
	Negative	-.244	-.372	-.375
Kolmogorov-Smirnov Z		1.464	1.464	2.233
Asymp. Sig. (2-tailed)		.028	.027	.000
Wilcoxon Signed Ranks Test		Cellular-Vehicle	Sensor-Cellular[b]	Sensor-Vehicle
Z		-3.582	-5.091	-3.990
Asymp. Sig. (2-tailed)		.000	.000	.000

6.2.7 I-695 Outer Loop @ US 1

This sensor location is matched to the TMC node 4539 in direction 0. The cellular data is available from 26 January to 3 February, while the sensor data is available from 0:50:00 on 26 January to 23:50:00 on 3 February 2006. On 27 January 28 probe vehicle speeds are available and 12 speeds are available on 3 February 2006. Figure 6-32 shows the value of speeds from all three data sources on 27 January and Figure 6-33 shows the ones on 3 February. Figures 6-34 and 6-35 focus on the time interval during a day in which all three types of data are available.

Speed deviations with respect to vehicle speed are presented in Figures 6-36 and 6-37. It is found that, at this location, the cellular speeds underestimate the vehicle speeds more than the sensor ones on average. Cellular speeds possess a larger absolute deviation. Table 6-13 summarizes the speed deviations at the TMC node 4539 in direction 0. Sensor data outperforms cellular data in this location.

Vehicle and sensor data are tested to be normally distributed, and t-test shows that they are closely related to each other. The difference between sensor-vehicle pair and sensor-cellular pair are statistically significant. Table 6-14 presents the statistics for I-695 outer loop at US 1.

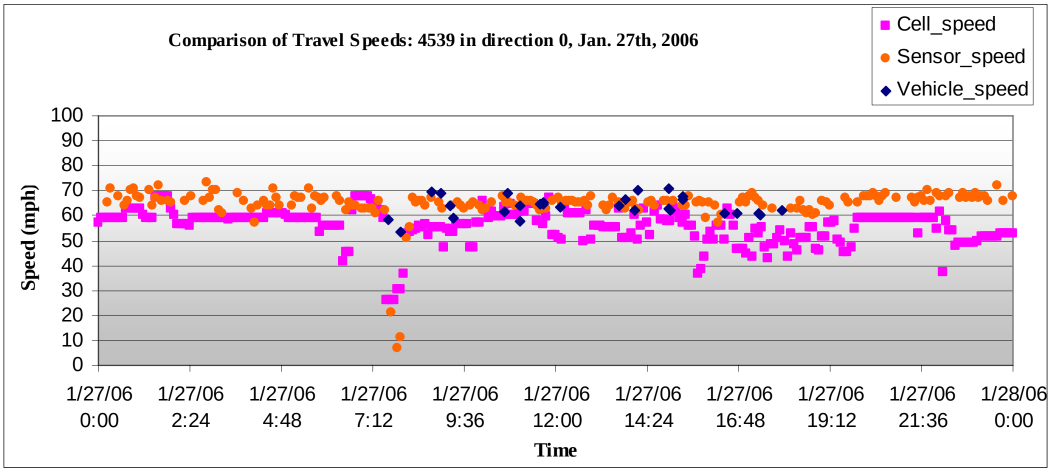

Figure 6-32
Comparison of Travel Speeds: I-695 Outer Loop @ US 1, 27 January 2006

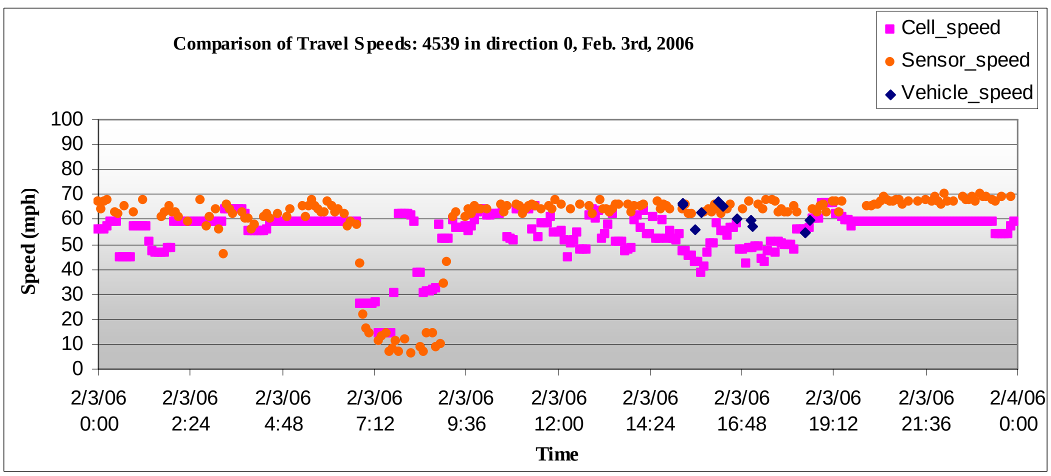

Figure 6-33
Comparison of Travel Speeds: I-695 Outer Loop @ US 1, 3 February 2006

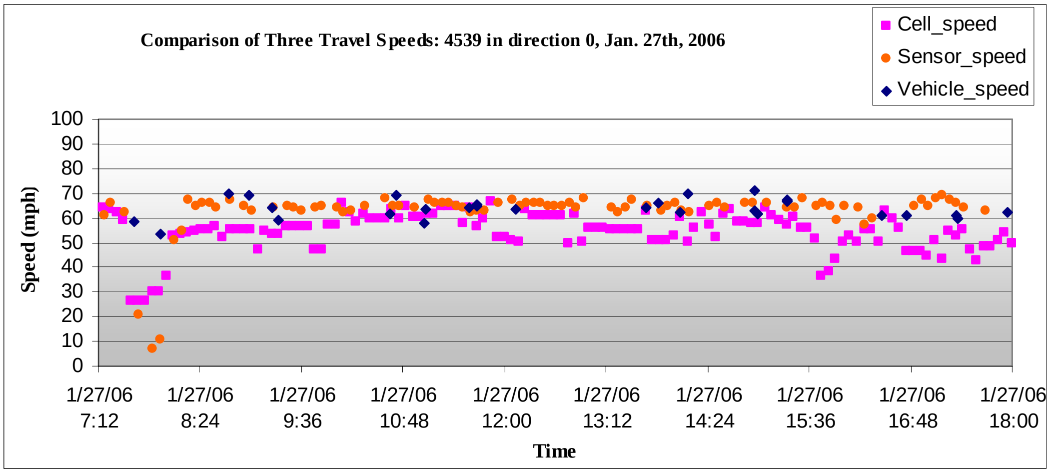

Figure 6-34
Comparison of Three Travel Speeds: I-695 Outer Loop @ US 1, 27 January 2006

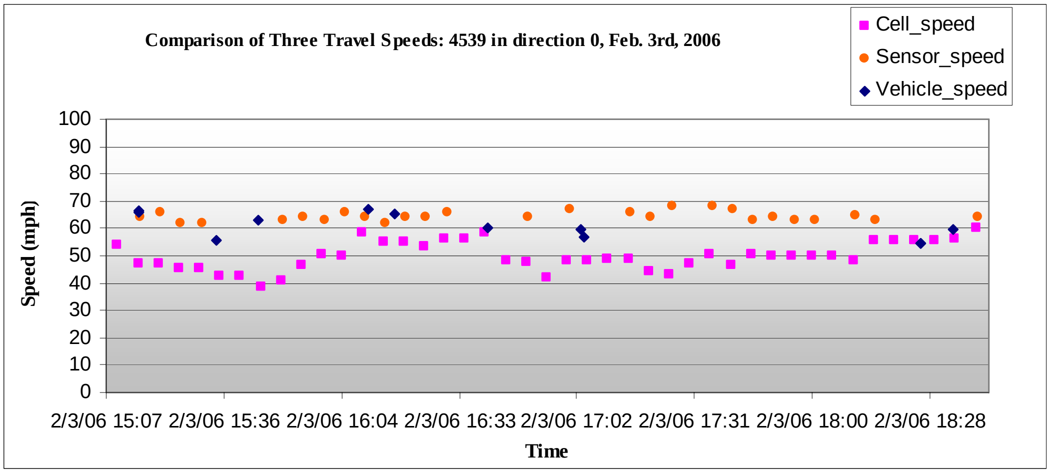

Figure 6-35
Comparison of Three Travel Speeds: I-695 Outer Loop @ US 1, 3 February 2006

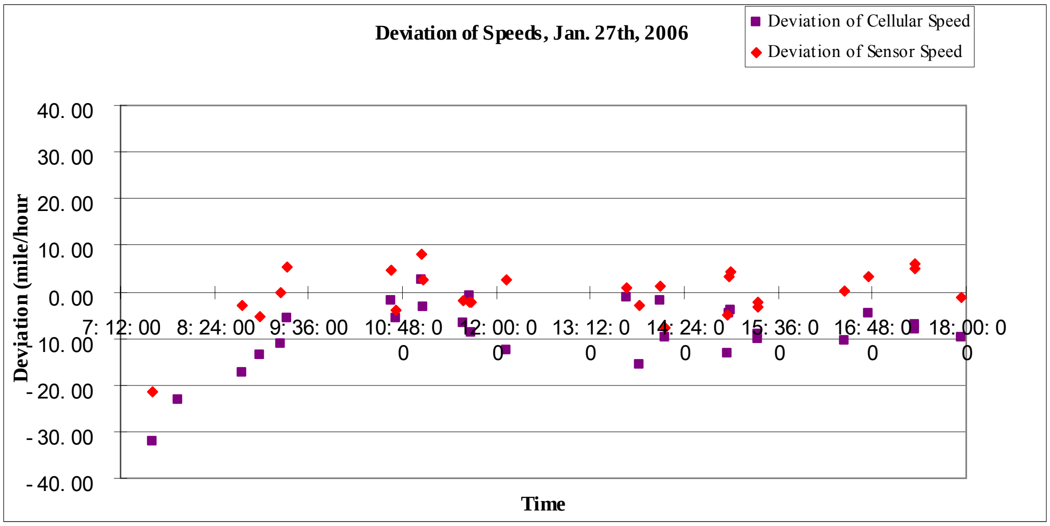

Figure 6-36
Speed Deviations: I-695 Outer Loop @ US 1, 27 January 2006

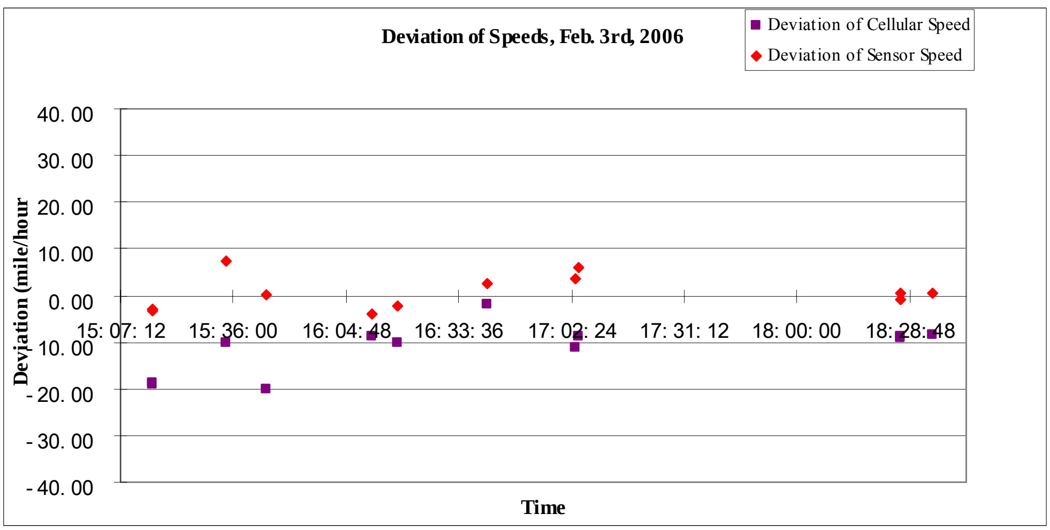

Figure 6-37
Speed Deviations: I-695 Outer Loop @ US 1, 3 February 2006

Table 6-13
Speed deviations: I-695 outer loop @ US 1

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-9.62	9.76	-1.26	5.03

6.2.8 I-695 Inner Loop @ US 1

This sensor location is matched to the TMC node 4539 in direction 1. Its corresponding location is I-695 inner loop at US 1. The cellular data is available from 26 January to 3 February while the sensor data is available from 00:55:00 on 26 January to 23:55:00 on 3 February 2006. There are 28 probe vehicle speeds available on 27 January and 12 speeds on the other day. Figures 6-38 and 6-39 show the value of speeds from all three data sources, while Figures 6-40 and 6-41 focus on the time interval during a day in which all three types of data are available. Based on the probe vehicle speeds, Figures 6-42 and 6-43 show the deviations of cellular-vehicular speeds and sensor-vehicular speeds. On average, both data underestimate the vehicle speed. Table 6-15 summarizes the speed deviations at the TMC node 4539 in direction 1. It is interesting to observe that on 27 January the sensor data overestimates the vehicular speed and the cellular data underestimates the vehicular speed, while on 3 February it is the opposite.

Table 6-16 presents the statistics of correlations between the data sources. It is observed that vehicle and cellular data follow normal distribution. Therefore paired t-test (parametric test) can be applied for cellular-vehicle pair. The results show that the these two groups of data have the same means. For the other two pairs of data, Wilcoxon test is used and the results confirm similarity in both cases.

6.2.9 I-695 Outer Loop @ I-95

This sensor location is matched to the TMC node 4540 in direction 0. The cellular data is available from 26 January to 3 February 2006, while the sensor data is available from 00:52:00 on 26 January to 23:57:00 on 31 January. On 27 January 28 probe vehicle speeds are available and 10 speeds are available for 3 February. Figures 6-44 and 6-45 show the value of speeds from all three data resources on 27 January and 3 February respectively. Figures 6-46 and 6-47 focus on the time interval during a day in which all three types of data are available. The deviations of cellular-vehicular speeds and sensor-vehicular speeds are shown in Figures 6-48 and 6-49.

Table 6-14
Statistics of travel speed comparison: I-695 Outer Loop @ US 1

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		40	40	40
Normal Parameters(a,b)	Mean	53.7698	62.0335	53.1858
	Std. Deviation	7.71992	9.43829	15.18727
Most Extreme Differences	Absolute	.158	.435	.375
	Positive	.119	.312	.206
	Negative	-.158	-.435	-.375
Kolmogorov-Smirnov Z		.337	.998	2.751
Asymp. Sig. (2-tailed)		1.000	.272	.000
Paired Sample Correlation Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		0.400(a)	-4.974(b)	-.726(b)
Asymp. Sig. (2-tailed)		.010	.000	.468

Paired sample t-Test
Wilcoxon Signed Ranks Test

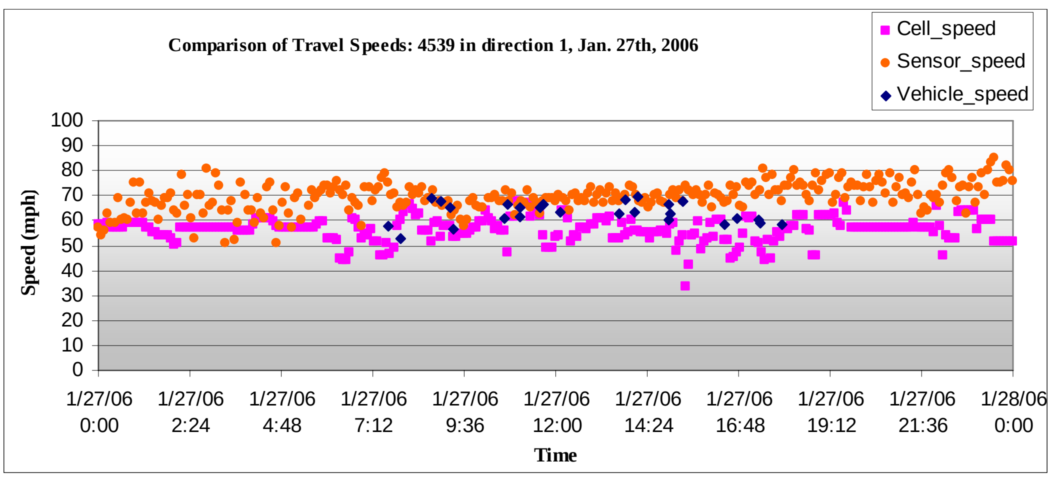

Figure 6-38
Comparison of Travel Speeds: I-695 Inner Loop @ US 1, 27 January 2006

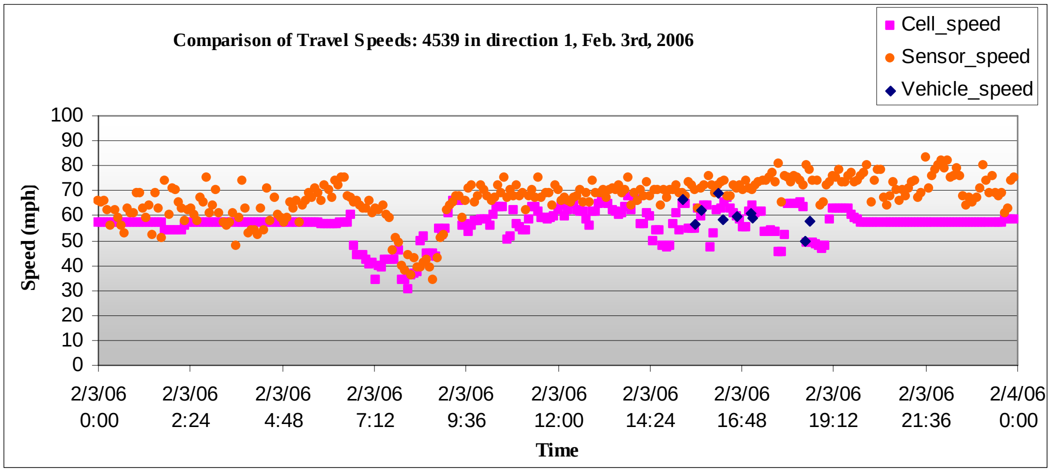

Figure 6-39
Comparison of Travel Speeds: I-695 Inner Loop @ US 1, 3 February 2006

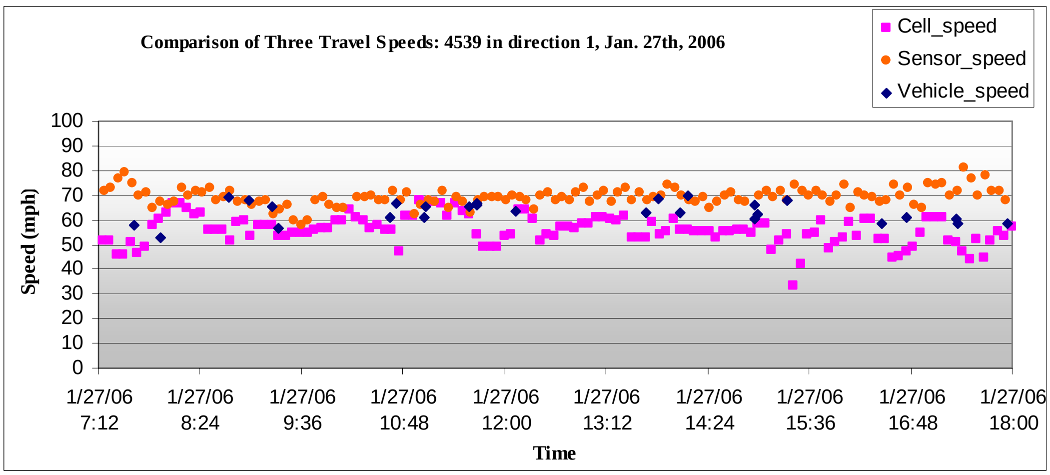

Figure 6-40
Comparison of Three Travel Speeds: I-695 Inner Loop @ US 1, 27 January 2006

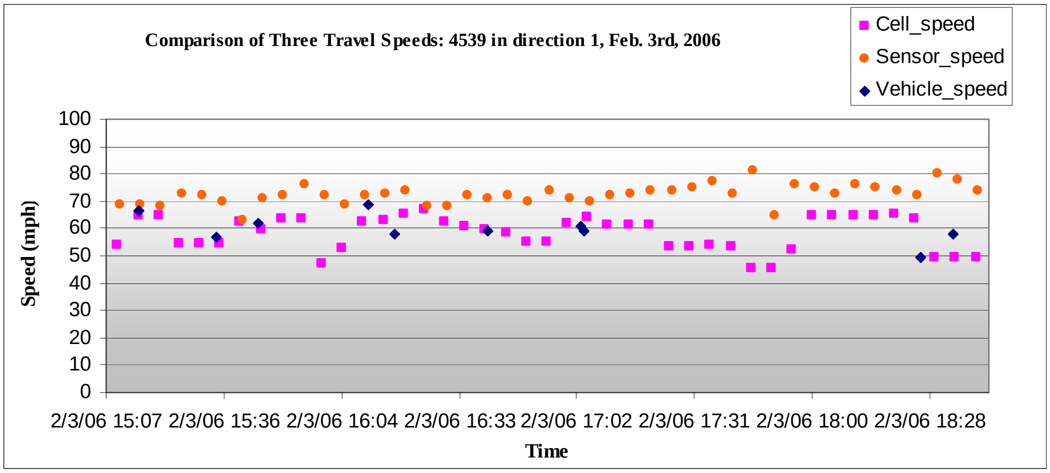

Figure 6-41
Comparison of Three Travel Speeds: I-695 Inner Loop @ US 1, 3 February 2006

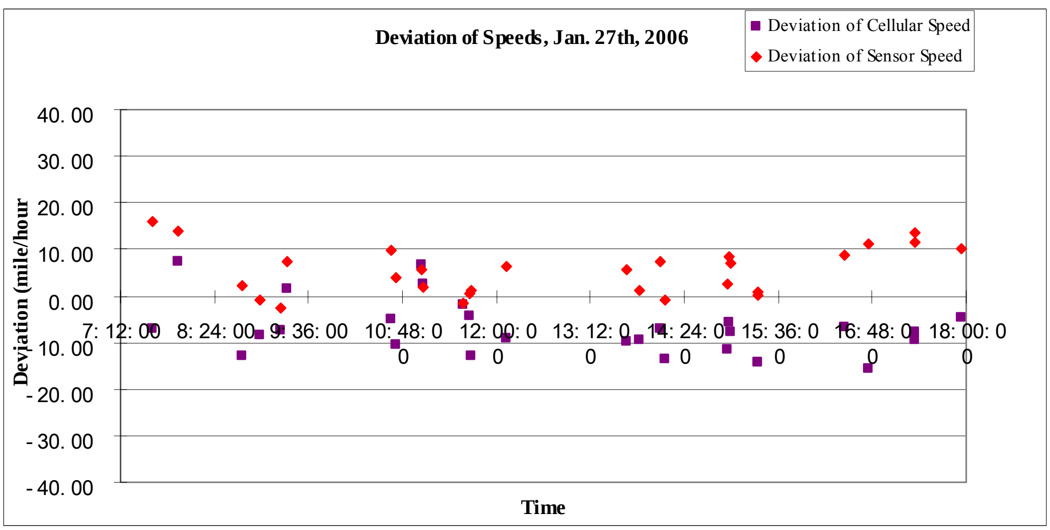

Figure 6-42
Speed Deviations: I-695 Inner Loop @ US 1, 27 January 2006

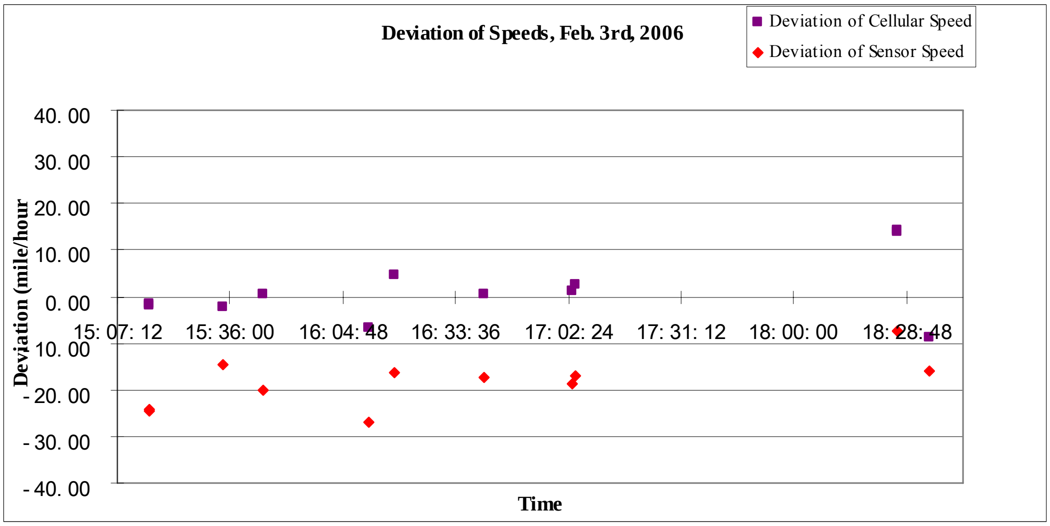

Figure 6-43
Speed Deviations: I-695 Inner Loop @ US 1, 3 February 2006

Table 6-15
Speed deviations: I-695 Inner Loop @ US 1

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-4.54	7.33	-1.49	9.35

Table 6-16
Statistics of travel speed comparison: I-695 Inner Loop @ US 1

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		40	40	40
Normal Parameters(a,b)	Mean	57.5922	60.6430	53.1858
	Std. Deviation	5.27669	12.54621	15.18727
Most Extreme Differences	Absolute	.136	.312	.375
	Positive	.136	.231	.206
	Negative	-.111	-.312	-.375
Kolmogorov-Smirnov Z		.810	.862	1.974
Asymp. Sig. (2-tailed)		.527	.447	.001
Paired Sample Correlation Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		0.008(a)	-.833(b)	-.155(b)
Asymp. Sig. (2-tailed)		.961	.405	.877

Paired sample t-Test
Wilcoxon Signed Ranks Test

For this location, it is observed that both cellular and sensor data underestimate the vehicle speeds. Table 6-17 summarizes the speed deviations at the TMC node 4540 in direction 0. Both data sources have large deviations from the vehicular travel speed.

One-Sample Kolmogorov-Smirnov test shows that all three data follow normal distribution, which allows us to apply a parametric standard t test. In all three cases, the null hypothesis is accepted, which means the three groups of travel speeds have the same mean. Table 6-18 presents the statistics of sample speed comparisons at I-695 outer loop at I-95.

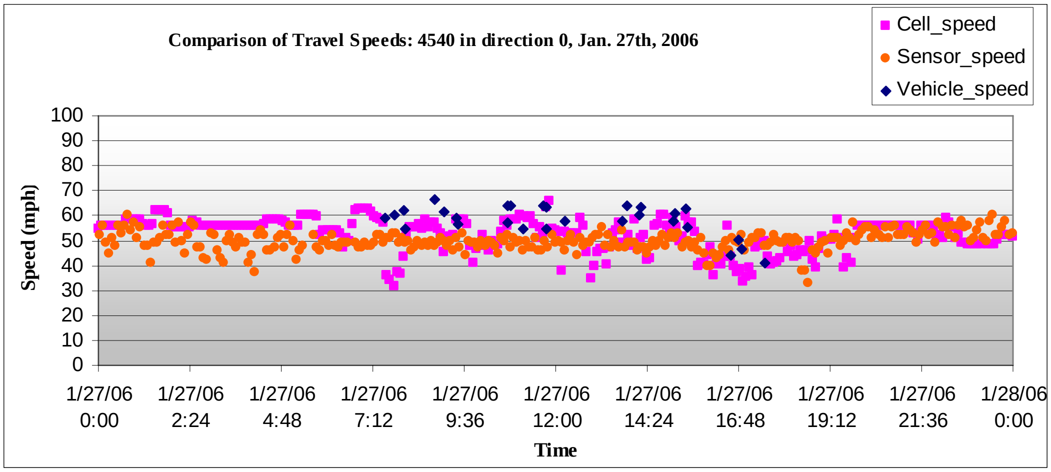

Figure 6-44
Comparison of Travel Speeds: I-695 Outer Loop @ I-95, 27 January 2006

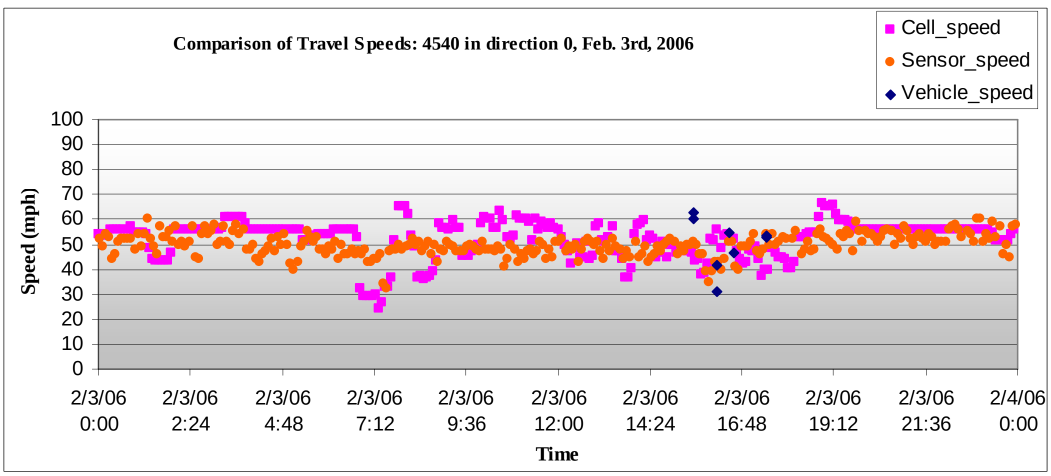

Figure 6-45
Comparison of Travel Speeds: I-695 Outer Loop @ I-95, 3 February 2006

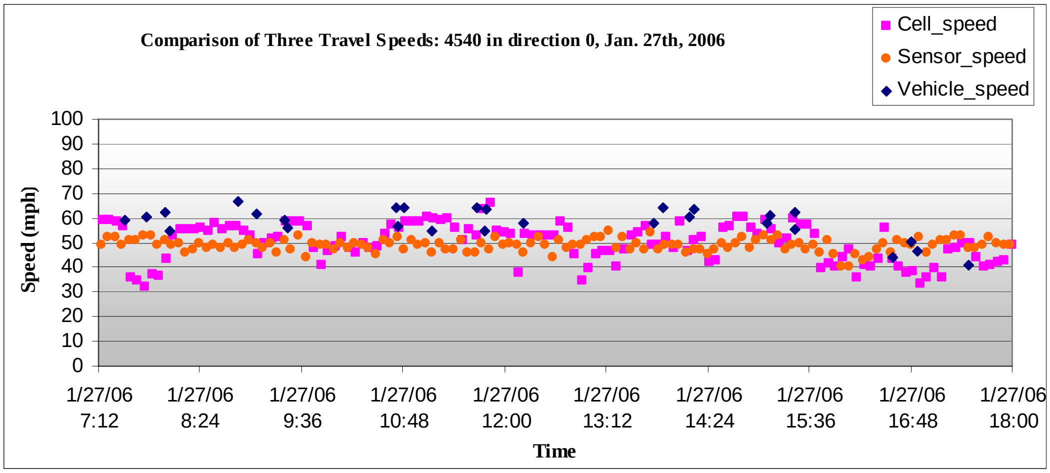

Figure 6-46
Comparison of Three Travel Speeds: I-695 Outer Loop @ I-95, 27 January 2006

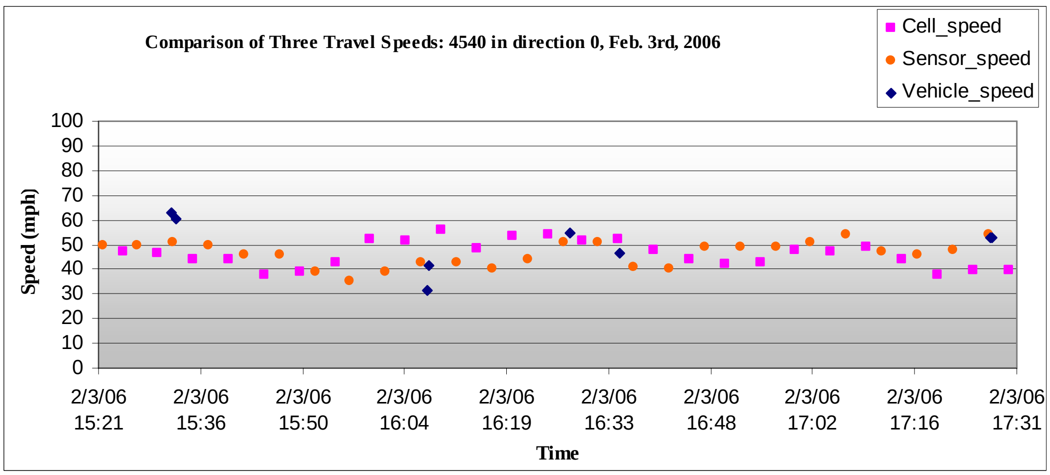

Figure 6-47
Comparison of Three Travel Speeds: I-695 Outer Loop @ I-95, 3 February 2006

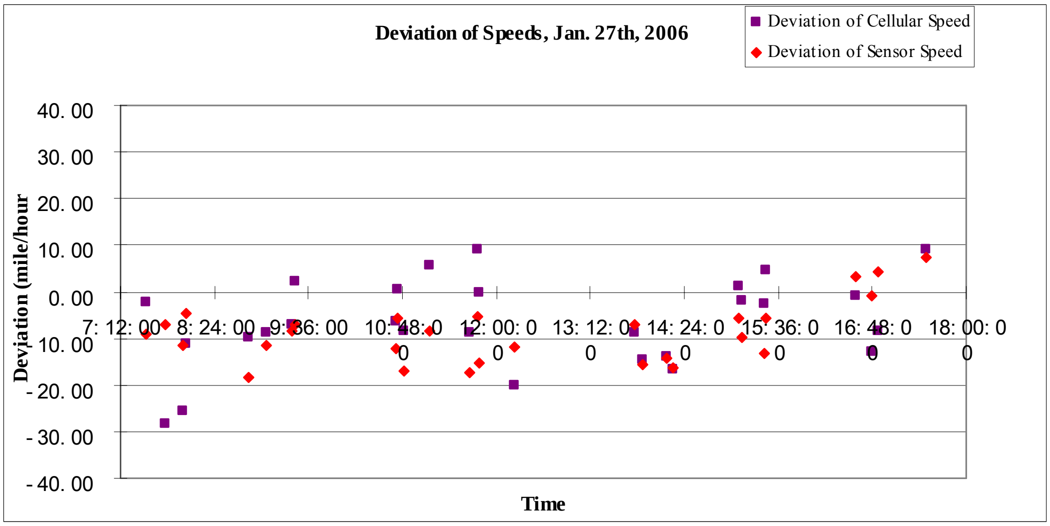

Figure 6-48
Speed Deviations: I-695 Outer Loop @ I-95, 27 January 2006

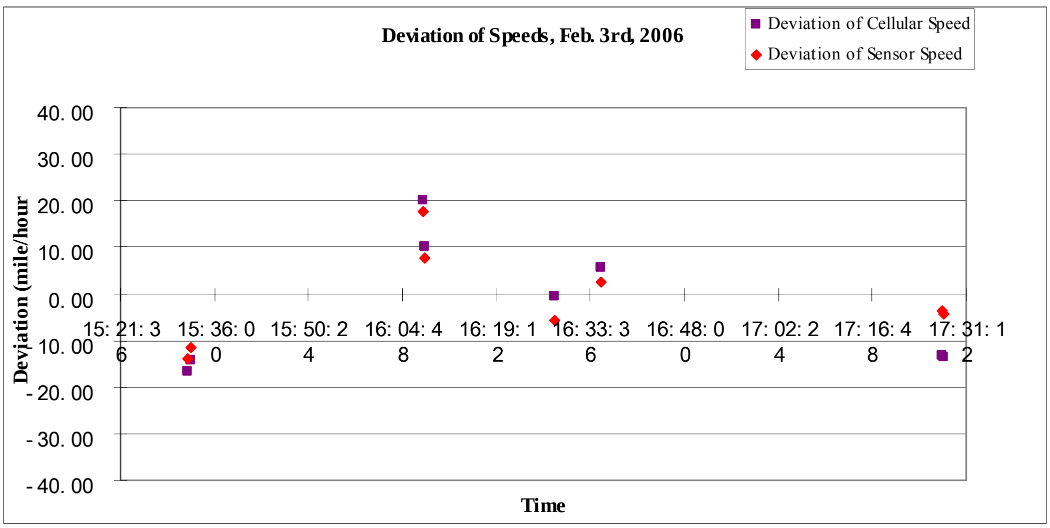

Figure 6-49
Speed Deviations: I-695 Outer Loop @ I-95, 3 February 2006

Table 6-17
Speed deviations: I-695 Outer Loop @ I-95

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-5.74	9.53	-7.11	7.47

Table 6-18
Statistics of travel speed comparison: I-695 Outer Loop @ I-95

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		36	36	36
Normal Parameters(a,b)	Mean	50.5611	49.1889	53.1858
	Std. Deviation	8.43400	1.70205	15.18727
Most Extreme Differences	Absolute	.108	.141	.375
	Positive	.100	.106	.206
	Negative	-.108	-.141	-.375
Kolmogorov-Smirnov Z		.931	.648	.846
Asymp. Sig. (2-tailed)		.351	.795	.472
Paired Sample T Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		.187	-.045	-.051
Asymp. Sig. (2-tailed)		.275	.793	.769

6.2.10 I-695 Inner Loop @ I-95

The last sensor location is matched to the TMC node 4540 in direction 1, and it corresponds to the location of I-695 inner loop at I-95. The cellular data is available from 26 January to 3 February while the sensor data is available from 00:52:00 on 26 January to 23:57:00 on 3 February 2006. Twenty-eight probe vehicle speeds are available on 27 January and twelve speeds are available on 3 February. Figures 6-50 and 6-51 demonstrate the speed distribution from all three data resources, while Figures 6-52 and 6-53 focus on the time period of the day in which all three data sources are available. Based on the probe vehicle speeds, Figures 6-54 and 6-55 show the deviations of cellular-vehicular speeds and sensor-vehicular speeds.

It is clear that cellular data underestimates while the sensor data overestimates the vehicle speeds. Table 6-19 summarizes the average speed deviations at the TMC node 4540 in direction 1. Cellular data has a smaller average deviation and smaller average absolute deviation.

It is found that all three data sources follow the normal distribution and Paired t-test is applied. Comparison of the vehicle-cellular pair shows the distributions of these two variables are significantly different. Table 6-20 summarizes the statistics of travel speed comparisons for I-695 inner loop at I-95. In this case, the cellular data outperforms the sensor data.

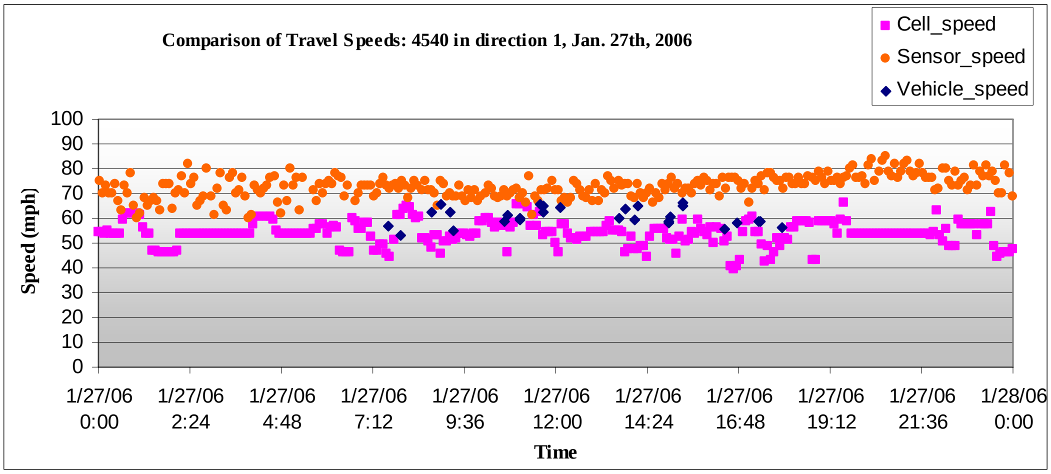

Figure 6-50
Comparison of Travel Speeds: I-695 Inner Loop @ I-95, 27 January 2006

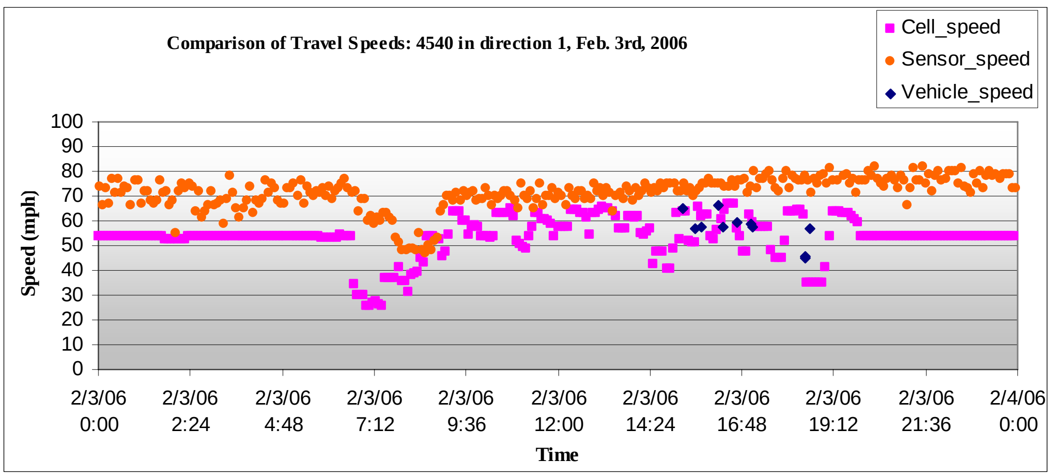

Figure 6-51
Comparison of Travel Speeds: I-695 Inner Loop @ I-95, 3 February 2006

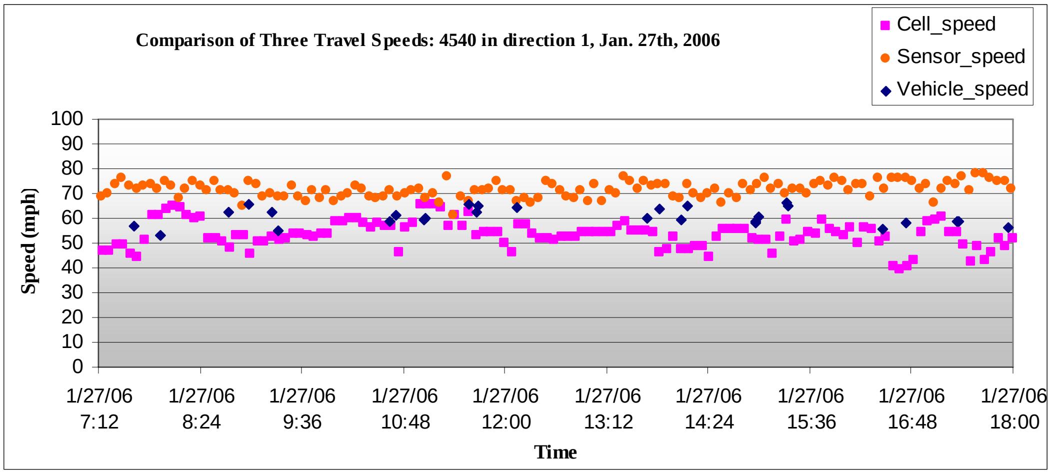

Figure 6-52
Comparison of Three Travel Speeds: I-695 Inner Loop @ I-95, 27 January 2006

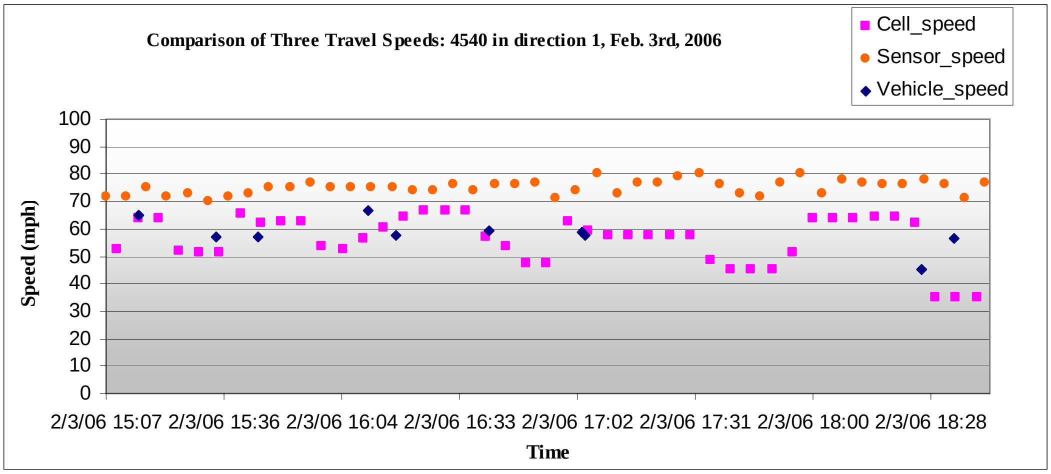

Figure 6-53
Comparison of Three Travel Speeds: I-695 Inner Loop @ I-95, 3 February 20

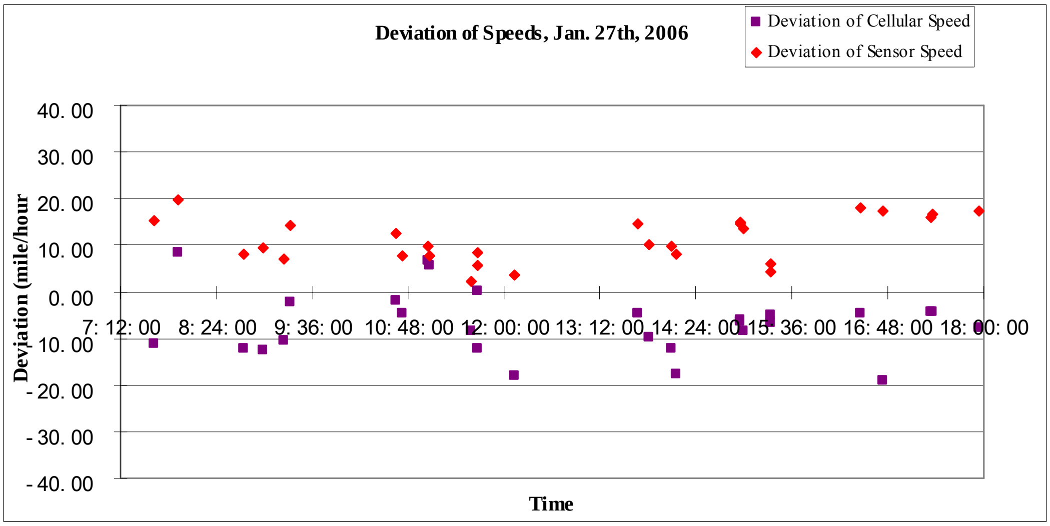

Figure 6-54
Speed Deviations: I-695 Inner Loop @ I-95, 27 January 2006

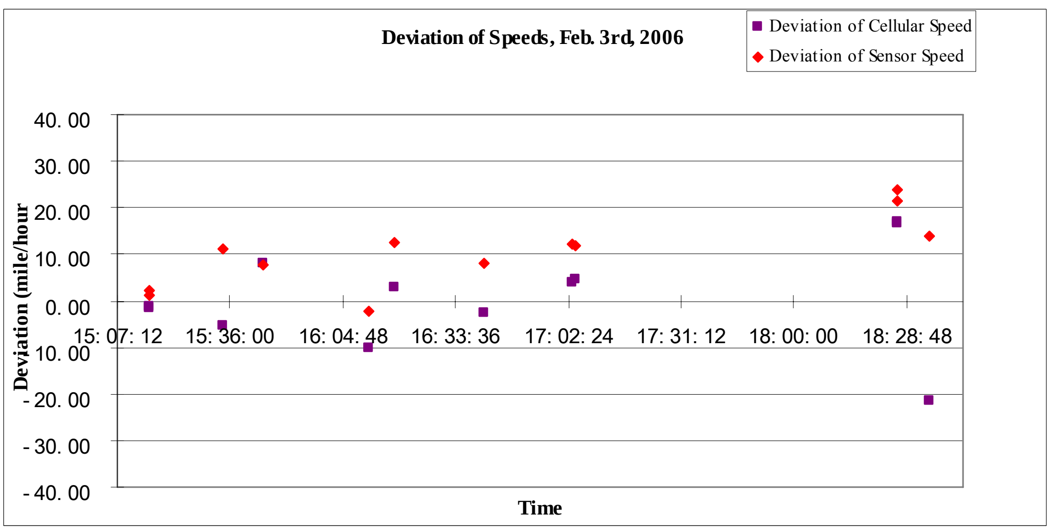

Figure 6-55
Speed Deviations: I-695 Inner Loop @ I-95, 3 February 2006

Table 6-19
Speed deviations: I-695 Inner Loop @ I-95

Cellular-Vehicular		Sensor-Vehicular	
Average Deviation	Absolute Deviation	Average Deviation	Absolute Deviation
-4.45	8.19	10.94	11.06

Table 6-20
Statistics of travel speed comparison: I-695 Inner Loop @ I-95

One-Sample Kolmogorov-Smirnov Test		Vehicle Data	Cellular Data	Sensor Data
N		40	40	40
Normal Parameters(a,b)	Mean	55.2486	70.6400	53.1858
	Std. Deviation	7.08863	2.98020	15.18727
Most Extreme Differences	Absolute	.087	.117	.375
	Positive	.066	.102	.206
	Negative	-.087	-.117	-.375
Kolmogorov-Smirnov Z		.723	.552	.738
Asymp. Sig. (2-tailed)		.672	.920	.648
Paired Sample T Test		Cellular-Vehicle	Sensor-Cellular	Sensor-Vehicle
Z		-.056	-.445	-.023
Asymp. Sig. (2-tailed)		.733	.004	.886

6.3 Summary

In this chapter, cellular and probe vehicle data were compared with data from fixed sensors. Data for ten sensor locations are available from the TMC database. Speed distribution from different data sources were shown for 27 January and 3 February 2006 (if available). Speed deviations with respect to the vehicle ones were calculated and statistical tests, parametric or non-parametric, were performed to test if the differences between any pairs of data are statistically significant. Findings and observations are summarized below.

Cellular data underestimates the vehicle speeds in 8 out of 10 cases, while sensor data overestimates in all 10 cases.

The average deviation of cellular data is better than the sensor data in 6 of 10 cases. In terms of absolute deviations, cellular data is worse than the sensor data in 6 out of 10 cases.

Cellular probe data outperforms sensor data in 3 out of 10 cases, while sensor data outperforms cellular probe data in 3 of 10 cases. In the other 4 cases, these two data source are tied with respect to their deviations.

The test results of cellular-vehicle pair shows significant difference in 4 out of 10 cases, while sensor-vehicle pair demonstrated significant difference in 6 out of 10 cases.

In general, the cellular data outperforms the sensor data, especially in capturing the overall picture of the variations in vehicular speeds.

7 Findings and Conclusion

The application of cellular probe vehicle data in ITS heavily depends on the quality of the data itself, especially in large scale real world operations. Cellular probe data quality evaluation has great importance in providing a solid base for future operations. In this evaluation project, vehicular probe data was used as the main comparison base although other data such as CHART detector data were also used.

An evaluation methodology was devised and the statistical tests were carefully selected based upon the nature of the data. Since the normality assumption was not satisfied in most cases for the variables of interest, various nonparametric tests were explored in this study. In the designed survey plan, the survey routes covered freeways and arterials, the survey areas included suburban area and downtown CBD area, and the survey times covered weekdays, weekend, morning peak hours, afternoon peak hours and non-peak hours. A total of 500 probe vehicle survey runs were conducted during the period 26 January to 3 February 2006. Due to budget limitations, there were 1 to 3 probe vehicles for a survey route in each time interval. The TMC road network was used and the trajectories of probe vehicles were matched to the TMC nodes for data quality evaluation. The cellular probe data of Baltimore metropolitan area spans from 26 January to 16 February 2006. The travel speed and travel time on each link are provided in every 5-minute interval. The vehicular probe data for each run is provided as a raw GPS file by MotionMaps. MotionMaps also matched the coordinates of probe vehicle locations to the TMC node numbers and provided the data as database files.

The cellular probe traffic data provided ideal spatial and temporal coverage for the entire Baltimore metropolitan road network during the test period of 21 days. There is great potential for use of this technology, especially when the data is augmented by data from other sources such as GPS, in future ITS applications such as traveler information system and real-time traffic control. The cellular probe data is superior to other methods in its coverage. The cellular probe data provides traffic speed and travel time information every 5 minutes. The accuracy of the travel time data obtained from the cellular probes on freeways is quite good (around 10% error terms) for immediate ITS application. For travel speed, the accuracy of the data obtained from cellular probes is acceptable (around 20% error terms). On arterials, the cellular probe data cannot provide accurate travel speed and travel time information at this time. There are several caveats relevant to the data performance measures. It is important to be aware of these issues in order to evaluate the cellular probe data quality.

The basic data quality measures are: average error, average absolute error, average error in percentage and average absolute error in percentage. Probe vehicle data is deems as the "Ground Truth" in the calculation of these measure. Theoretically, probe vehicles can provide average travel speed when drivers strictly follow floating car method. In practice, when the traffic is congested on road, it is difficult for a driver to pass one car for every car that passes him or her. Another problem is that when there are many lanes on a freeway segment, it is difficult for the driver to observe the movements of all vehicles in every lane. Similar problems occur when traffic lights are installed on streets. Statistical sampling methods suggest that roughly 4% to 5% of probe vehicle penetration can provide a good estimate of travel times. Based on the statistical concept of allowable standard errors, the study showed that approximately 150 to 160 observations per hour would be necessary for reliable speed estimates. As some observations will likely need to be discarded as problematic, a larger number of observations will be needed in a deployed system. In our study, the number of probe vehicles is far less than the recommended values. Therefore the vehicle probe data that are used in this study may not accurately represent the "Ground Truth".
Besides the "Ground Truth" issue, default values are another issue that may greatly influence the data quality measures. According to ITIS Holdings, on some routes where there is no real time data available, default value (usually, the speed limit) are provided as cellular probe speed. These default values sometimes are far from the vehicular speed and thus generate large errors and degrade the data quality measures. The deficiency of cellular data may be caused by low penetration of cellular phone or limitations of cellular carrier. This issue might be solved by contracting with a cellular carrier with high market penetration rate.

The outlier impact can influence the quality measures as well, especially for link speed comparison. This impact is caused by the extremely slow speeds. The vehicular data is used as the "Ground Truth" in the evaluation. When calculating the percentage of errors, probe vehicle speeds are used in the denominator. Huge percentage of errors will be generated when the absolute error between the cellular speed and vehicular speed is relatively large. As an example, when the vehicular speed is 5 miles per hour and the cellular speed is 45 mile per hour, the absolute error is 800 percent. These outlier data points contribute to increasing the average percentage of absolute error. Therefore, it is important to pay attention to the histogram analysis of the data when drawing conclusions from the percentage of average and absolute errors that are presented in this report.

These issues can bring large variability to the data and greatly degrade data quality measures. The conclusions herein are made on relevance and similarity rather than correctness of traffic information on road segments. Clarifications of these issues can help readers to understand the data reported in this evaluation project.

The evaluation of travel speeds was performed on links because the link length discrepancies can have major impacts on the link travel time comparison. As shown in

Table 4-1, the average absolute error is around 20% of the vehicular link length. Though the majority of paired link travel speeds have a deviation of 10 to 20 miles per hour, the statistical test results indicate that there is no significant difference between cellular probe speeds and vehicular probe speeds on major freeways. That means, on freeways, cellular probe data can provide reasonable estimations of link travel speeds. On arterials, however, the speeds from cellular probe data are significantly larger than those from probe vehicle data. As an example, on MD Route 40 which runs through the CBD area in Baltimore, the average cellular speed overestimated the average vehicular speed by 14.37 miles per hour. The average absolute deviation is 17.39 miles per hour, which is more than 60% of the probe vehicle travel speeds. The comparison of average link speeds on each route and the corresponding histogram analysis of the average absolute error on each route are illustrated in Figures 4-2 and Figure 4-4. The analysis of link speeds by time of day shows that the differences between these two types of data are larger during the peak hours. Figures 4-5 and 4-6 show the comparison of average link speeds and average errors by time of day. This finding leads to the comparison of link speed by Level of Service (LOS). The analysis shows that the deviations are highly correlated to the LOS. When the LOS degrades, the data quality measures degrade correspondingly. Figures 4-11 and 4-12 illustrate the results of the histogram analyses of average absolute errors by the LOS.

For paths composed of multiple links, travel time evaluation was conducted. The path length analysis result shown in Table 5-1 indicates that the majority of paths have small length discrepancy less than 10%. Travel time evaluation shows the same trend in link travel speed evaluation. That is, cellular probe data provides good estimations of path travel times on freeways. The deviations between cellular path travel times and vehicular path travel times range from 7% to 13%. The summary of the travel time comparison on each route is shown in Table 5-4. The travel times obtained from the cellular probe data are significantly shorter than those obtained from the probe vehicle data on arterials. The analysis shows that the cellular probe travel times consistently underestimated the vehicular travel times and the average absolute deviation is around 30% to 60% of vehicular path travel times on arterials. Figures 5-1 to 5-3 demonstrate the comparison of average travel times, average errors and histogram analysis results of average error on each route.

The comparison among cellular probe data, sensor data and vehicular probe data were performed based on the data collected from 10 sensor location on Interstate 695. The results show that all three technologies produced individual data points that were away from the others. Besides the temporal and special coverage advantage, the cellular probe data performed better in capturing the average travel speed over a time interval and the deviations were more evenly spread across the X axis, while the sensor data has a smaller average absolute deviation. The statistical tests show that the distribution of cellular speeds is similar to that of vehicular speeds in 6 out of 10

cases, and the distribution of sensor speeds is similar to vehicular one in 4 out of 10 cases.
Cellular phone tracking, at least the one implementation we examined, shows promise but further study is needed to determine its ability to produce quantifiable travel time data particularly for arterials. On the positive side, it promises very large numbers of probes and better spatial and temporal coverage then conventional loop detectors. However, the evaluation results indicate that there is a direct relationship between the road type and quality of estimated traffic information. On most freeways, the evaluation results are consistently good, on arterials in which traffic signals exist, the cellular probe data tends to over-predict the travel speeds and under-predict the travel times significantly. Another issue, which is particularly problematic for cellular probe data, is the effects of congestion on data quality relationship. When the level of service is low, the cellular data has larger deviations with the probe vehicle data and the average absolute deviation is around 40 to 50 percent.

For cellular phone tracking systems, further examination of the nature of the errors, particularly an additional comparison on arterials, is needed before any error correction algorithms could be developed. Additional information, such as signal phase settings could be useful in analyzing cellular probe data quality and should be provided if available.

References

[1] Yim,Y.B Youngbin and Randall Cayford (2001), "Investigation of Vehicles as Probes Using Global Positioning System and Cellular Phone Tracking: Field Operational Test." Working Paper.
[2] http://www.infoworld.com/article/06/04/07/77227_HNcellphonerecord_1.html?TELEPHONY
[3] "Defining and Measuring Traffic Data Quality." Office of Policy, Federal Freeway Administration (2003), Washington, DC.
[4] ITS America and U.S. Department of Transportation. "Closing the Data Gap: Guidelines for Quality Advanced Traveler Information System (ATIS) Data." Version 1.0, September 2000.
[5] Tarnoff, P.J (2002),"Getting to the INFOstructure." White Paper prepared for the TRB Roadway INFOstructure Conference.
[6] Yim, Youngbin, Adib Kanafani, Jean-Luc Ygnace (1991) "Expanding usage of cellular phones: userprofile and transportation issues." PATH Research Report, UCB-IS-PRR-91-19.
[7]Smith, Brian L., Han Zhang, Mike Fontaine and Matt Green (2001), "Cell Phone Probes as an ATMS Tool." Project report for the Center for ITS Implementation Research (ITS).
[8] http://www.dot.state.fl.us/TrafficOperations/ITS/Projects_Deploy/Newsletters/2006/04_2006/04_2006_page1.htm#Cell_Phone_Location_Systems_State_of_the_Practice_
[9] "Missouri DOT CFVD Baltimore System Evaluation v1.2." Internal tech. report. IT IS Holdings, 2006.
[10] Steven K. Thompson (2002), "Wiley Series in Probability and Statistics." Second Edition, Wiley-Interscience.
[11] Sachs, Lothar (1982), "Applied Statistics: a handbook of techniques (Springer series in statistics)" Second Edition, Springer-Verlag.
[12] Kleijnen, Jack (1999), "Validation of Models: Statistical techniques and data availability", Proceedings of the 1999 Winter Simulation Conference, P. A. Farrington eds.
[13] J. D. Gibbons, *Nonparametric methods for quantitative analysis*. New York: Holt, Rinehart and Winston, 1976.
[14] J. D. Gibbons and S. Chakraborti, *Nonparametric statistical inference*. New York: Marcel Dekker, 1992.
[15] S. Wegenkittl, "Entropy estimators and serial tests for ergotic chains", *IEEE Trans. Inform. Theory*, vol 47, no. 6, pp. 2480-2489, Sept. 2001.