Infrared Remote Sensing
Of On-Road
Motor Vehicle Emissions
In Washington State

Appendices

SANTA BARBARA RESEARCH CENTER

A Subsidiary of Hughes Aircraft Company

75 COROMAR DRIVE, GOLETA, CALIFORNIA 93117

1996 REMOTE SENSING STUDY
FINAL REPORT

Prepared For
WASHINGTON STATE DEPARTMENT OF ECOLOGY
300 Desmond Drive
Lacey, WA 98503

Contract No. C9700021

Publications Reference No. 97-0018

January 1997

HUGHES
AIRCRAFT

21 January 1996
In Reply Refer To: 97- 01 1:LL78

Washington State Department of Ecology
300 Desmond Drive
Lacey, WA 98503

Attention: Kim Wager

Subject: Contract No. C9700021 Remote Sensing Study;
Final Report

Upon delivery of this document, we are pleased to report that Santa Barbara Research Center (SBRC) has successfully met all project goals of the Washington Department of Ecology (WDOE) 1996 Remote Sensing Study. Data collection was completed on 30 September 1996. Since contract award *on* 9 July 1996, SBRC accomplished the following:

1) Designed a database to track each of the parameters required by WDOE;
2) Modified the Remote Emission Sensor systems for unique study requirements;
3) Met with WDOE personnel on 9 and 13 August 1996 for the program kickoff and final planning meetings;
4) Conducted site selection and permitting activities;
5) Collected emissions, license and vehicle speed data in the Pierce, King and Snohomish counties from 26 August through 30 September 1996;
6) Delivered all required data collected during this study;
7) Submitted a final report summarizing the study results.

The following report is a summary of the activities and results of this study. The information on data statistics is described or categorized as "Valid" and "Invalid" data. The term "Valid" refers to all data that has met the requirements established by the WDOE for use in this study. The term "Invalid" refers to all data that falls outside these same requirements.

If you have any questions regarding the above, they may be directed to the undersigned at (805) 562-4398.

SANTA BARBARA RESEARCH CENTER

Sylvia M. Salazar
Sr. Contracts Negotiator

snis \ 1178frpt

Santa Barbara Research Center
75 Coromar Drive
Goleta, CA 93117(805)
968-3511

CONTENTS

Section

1.0 REPORT SUMMARY
  1.1 Site Selection
  1.2 Permitting Activities
  1.3 Data Collection
  1.4 Data Statistics

2.0 DATABASE
  2.1 Database Design
  2.2 Database Structure
  2.3 License Plate Data

3.0 MOBILE LAB AND SUPPORTING EQUIPMENT
4.0 SITE SELECTION
5.0 PERMITTING
6.0 DATA COLLECTION
7.0 CLOSING COMMENTS AND OBSERVATIONS

Table

2-1 Output Parent Database Structure
2-2 Supporting Databases
2-3 Code legends for License Plate Fields
2-4 Other Codes Generated By System
7-1 Comparison Table of Data Calculated from the “Valid” Records to “Invalid Records

1.0 REPORT SUMMARY

The following is a summary of the activities and results of the Washington Remote Sensing Study of September 1996.

1.1 Site Selection

- Site selection and permitting were completed within a six-day period.

- Actual site selection was continuous over the five weeks of the study. Changes were made as required to meet the study targets.

- A total of 62 sites were identified as possible collection locations. This list was condensed to approximately 20 sites considered to have the best potential for the expected results.

- A total of six sites were actually used for this study.

1.2 Permitting Activities

Permitting the site locations for this study involved contact with four state agencies and five departments:

- Washington Dept. of Ecology:
  Kerry Swayne
- Washington Dept. of Transportation - Traffic Engineering Div:
  Elizabeth Church/Carol Larson
- Washington Dept. of Transportation - Northwest Region:
  Roger Stienert
- Washington Dept. of Transportation - Pierce Co./Southern Region:
  Frank Newboles
- Washington State Police:
  Lt. Hurlbut

No formal permits were required by any party. Weekly contact was required with the above agencies and departments advising them of site locations and time schedules. Faxes were sent on a regular basis to all the involved agencies.

1.3 Data Collection

- Collection occurred on 26 of 36 days at six sites within the three counties.
  Roadside data collection started on 26 August 1996 and ended on 30 September 1996.

- There were no weekend collection days.

- Normal collection times were between 7 AM and 3 PM. Daily times varied due to weather and traffic conditions.
  No major equipment failures occurred during the course of this study.

- No equipment was lost, stolen, or damaged.

- No abnormal traffic situations were observed due to the presence of our equipment.
  - No concerns over safety or traffic problems were reported by any of the involved agencies during this study.
  - Traffic in the three-county area was well mannered and did not create a hazard for our equipment or personnel.
  - Most drivers were courteous and appeared to be unaffected by the presence of our equipment.
- Weather did not appear to play an important roll in the data collection activities during this study.
  - Weather conditions for most of this study can be summarized as: clear to light overcast ski6s, normal temperatures ranging from upper 60's to mid 70's, with very light to non-existing wind conditions.
  - Rain prevented collection for most of one day.
  - Rain or wet road conditions caused a loss of several hours of collection on several other days.

1.4 Data Statistics

- A total of 201,581 records were collected over the 26-day collection period.
  The lowest volume of records collected was 2,233.
- The highest volume of records collected was 13,185.
- The daily average collected was 7,753.
- Of the total 201,581 records collected, 73,498 records were lost due to being registered outside the non-attainment area.
  - The average loss per day due to out-of-area registration was 2,827 or 36.4% of the average daily collection.

- Of the total 201,581 records collected, 22,013 records were lost due to invalid CO readings.

  The average daily loss due to invalid CO was 846 or 10.91 % of the total records collected.
  The total number of records lost due to invalid speed readings was 5,840.
  The average daily loss due to invalid speed readings was 225 or 2.89% of the total records collected.

- The total number of valid records collected in the 26-day collection period was 64,028.

  The average daily collection of valid records was 2,463 or 31.38% of the total daily collection.

- The total number of three or more measurements (hits) on the same vehicle was 4,539.

The total cost per valid record was $0.76.

2.0 DATABASE

The database structure and report functions for this study were based on the following criteria:

- Real time emissions data along with an identifying vehicle number stored in ASCIII format.

    1. Emissions, license plate, and speed-acceleration data (up to 13 parameters) obtained on each vehicle
    2. Plume strength ratios of CO, HC, and NOx toC02
    3. Effective tailpipe concentrations of CO, HC, NOx, andC02
    4. Data validity codes for CO, HC, and NOx
    5. License plate data with linking record number to emissions data

Weather information for each collection day.

1. Information included general conditions such as clear, sunny, overcast, and rain
2. Temperature readings taken during each collection period
3. Wind speed and direction, if present
4. Humidity

Site information for each collection period.

1. Location, date, number of vehicles, and time of collection

2.1 Database Design

The database designed for this study was adapted from the database created for the ongoing remote emissions collection program in Phoenix, AZ. Many of the elements were already in place since the Phoenix database comprised all of the basic emissions, license plate, and speed reading capabilities required for the Washington study. The real-time emissions data along with an identifying vehicle number is stored in the required ASCH format. Some additional tables and specialized reporting functions were added or modified to meet our needs of reporting in area vehicles, multiple hits per vehicle, weather, and site information. This task was completed by our Phoenix based Data Administrator, Ms. Susan Reed. Ms. Reed made the changes to the original database and tested the final version using previously collected data.

While testing these newly developed tables and report calculations, Ms. Reed found that some of the elements were not working as planned when the software was used in the mobile lab's computer environment. Further design modifications and testing proved the database to be functional; however, the added time to develop and test the new software forced a change in our original strategy. Our original plan had called for the raw data to be merged and processed onsite by the operator collecting the data. In this way, we would have a fast turn around of the daily collection results and could make adjustments in our activities as needed. With the delay in testing and debugging the new software, it was necessary to process the data at our Phoenix facility rather than on-site. The data transfer process was accomplished by modem connection directly from the mobile lab to our Phoenix server.

2.2 Database Structure

The structure of the output parent database for this study is shown in Table 2-1.

Table 2-1. Output Parent Database Structure*

	Field Name	Description
1	VEH NUM	Computer generated vehicle tracking number
2	DATE	Date of data collection
3	TIME	Time stamp of each record created
4	COPERCENT	CO percent measured
5	C02PERCENT	C02 percent measured
6	HCPERCENT	HC percent measured
7	SENSOR	Serial number of sensor collecting data
8	LPLATE	License plate information
9	STATE	Two letter state code
10	SLOPECO	Calculated slope of CO/CO2
11	SLOPEHC	Calculated slope of HC/CO2
12	VELL	First speed reading entering bigger zone
13	VEL2	Second speed reading existing trigger zone
14	DELTA	Calculated acceleration factor
15	NOPERCENT	NO percent measured
16	SLOPENO	Calculated slope of NO/H20
17	HE	Always 1, not used
18	VAN	Mobile Lab unit number, always 6 for this study
19	RECID	Record number
20	LOCATION	Site location information

*Database Name: EHISTORY.DBF - Borland Dbase engine

To be entered in the parent database, a record must meet a particular criteria. A valid record in the EHISTORY database is comprised of the following:

- Carbon Monoxide (CO) is not a value of 999.
- Carbon Dioxide (C02) is not a value of 999.
- A license plate number listed in the VALD:)PLT.DBF database, not a manually entered code.
- VEL I reading other than 99.
- VEL 2 reading other than 99.
- DELTA (Acceleration) other than 999.

In addition to EHISTORY.DBF, there are supporting databases to store additional information (refer to Table 2-2). The STATS.DBF and FUTCOUNT.DBF supporting databases contain calculated information based on data found in the El-HSTORY.DBF. These reports, contained in Appendix A and B, respectively, may be very beneficial in determining the overall effectiveness of the study.

Table 2-2. Supporting Databases

Database Name	Description
HITCOUNT.DBF	Running accumulation of total records, valid records, invalid percents, etc.
SLOOKUP.DBF	List of all sites considered for this study
STATS.DBF	Daily statistics recording the valid percentage of total records, invalid percentages of CO, VEL 1, VEL2, DELTA SPEED, and number of invalid license plate reads by letter code
VALIDPLT.DBF	List of all "In Area" license plates. List supplied by WDOE for validating records
WEATHER.DBF	Includes date, time, temperature, humidity, and wind speed and direction information recorded on each day of collection
WASHSITES.DBF	Record of each collection day location, site code number, date, time, mobile lab unit number, and operator name
BAT****.DBF	Raw data for each collection day. Includes valid and invalid data

2.3 License Plate Data

During typical data collection, license plate data is obtained by using a separate Automatic License Plate Reader (ALPR) computer linked to the emissions computer. The ALPR system captures a high resolution black and white image of the rear of a passing vehicle, searches the scene for a license plate image, and stores the image as a Target Image File Format (TIFF). Where possible, it also attempts to interpret the image into a license plate number. The TIFF images are linked to the emissions data by a corresponding unique record number created in the emissions computer. As each vehicle license plate is captured and interpreted, the ALPR sends the plate information to the emissions computer where it is linked to the proper vehicle emissions record. The result is an ASCII file containing the alphanumeric representation of both the emissions and license plate information. In some cases, TIFF images captured by the ALPR are not able to be immediately "recognized" by the system and are later manually "truthed" by means of an application program after data collection has ended. These records are merged into the emissions records with the vehicle number as the correlation variable.

The process of reading the captured image requires that the ALPR system recognize a license plate by means of a pre-defined database of license plate formats. Each state has a unique format and/or font set, that must be created and stored in the ALPR system for this process to function correctly. Due to time and budget considerations, a font set for Washington was not created. All license plate data was manually entered after data collection ended.

Although this process increased the time expended to complete the data set, the end result was functionally the same. To expedite the process of manually entering the 200,000 license plates collected during this study, we enlisted the help of our van operators assigned to the Phoenix facility. The files containing the license plate images were transferred to Phoenix where they were "truthed" and "merged" with the emissions data before processing the database.

Code legends for license plate fields, as well as other codes generated by the system, are identified in Tables 2-3 and 2-4.

Table 2-3. Code legends for License Plate Fields

NO READ	Code generated by the ALPR system. Denotes a trigger that occurred when there was no image for the system to analyze or attempt to read.
NO-FIND	Code generated by the ALPR system. Denotes a trigger that occurred when there was an image for the system to analyze but the system could not find anything it could recognize as a license plate
NOPLATE	Code generated by Emissions computer system to indicate a trigger has occurred but the system could not recognize, or did not receive data from the ALPR system within the proper time window. Occurs more frequently when traffic is heavy and vehicles are close together.
D	Code generated by operator when truthing license plate file. Denotes a Dealer plate.
E	Code generated by operator when truthing license plate file. Denotes an empty license plate holder. The image is of the area where a license plate should be but no plate is available.
F	Code generated by operator when truthing license plate file. Denotes a field out of view or partial plate image was captured by the ALPR system. Reflected light, improperly mounted license plate, or partially covered by soil or foreign object. Not all the letter or numbers are readable.
H	Code generated by operator when truthing license plate file. Denotes a hidden license plate. View is obstructed by some type of object.
0	Code generated by operator when truthing license plate file. Denotes Out of State or US Government vehicle plate.
P	Code generated by the operator when truthing license plate file. Denotes a paper or temporary license plate.
T	Code generated by operator when truthing license plate file. Denotes truck or trailer license plate.
U	Code generated by operator when truthing license plate file. Denotes an unreadable plate. Image is present but cannot be reliably read. Image blurred or bad light conditions are frequent causes of unreadable plates.
X	Code generated by operator when truthing license plate file. Denotes a known trigger for a calibration record.

Table 24. Other Codes Generated By System*

99999	Denotes an invalid measurement. The system will automatically enter a value of 99999 in the field for HC or NOx if the result does not translate to a valid measurement when compared to the valid/invalid software decision process incorporated in the system.
999.00	Denotes an invalid measurement. Same as above for the CO and CO, fields.

*Note - "0"or "0.0" is a valid number. The system does not substitute this number for any reason. A valid measurement of less than '7 may be rounded up to 0, but in no case will a 0 be used to represent a "Invalid Measurement condition.

3.0 MOBILE LAB AND SUPPORTING EQUIPMENT

To conduct the Washington study, Mobile Lab Unit No. 6 was outfitted with unique additional equipment required to meet the study goals. The basic system in this particular unit consists of an emissions sensor and emissions processing computer system, standard license plate reader, an upgraded Automatic License Plate Reader computer and camera system (ALPR), and a third truthing/data storage computer. A cellular phone, CO monitor, and color video monitor are also included in this unit's normal configuration. This unit's normal configuration also includes network capabilities to all three onboard computers with a network hub outlet for downloading data to a remote central server computer. A modem installed in the truthing/data storage computer was also added. The modem capabilities allow for data transfer to any location.

To meet the needs of the Washington study, it was necessary to add a Speed and Acceleration system and interface to the emissions data computer system, and install a Davis Weather Station base unit and related data collecting equipment.

The Speed system consist of two pressure activated sensor tubes, air operated switch interface, and a stand-alone micro processor unit developed by KGI to receive the speed data and prepare it for transfer to the emissions data computer.

The Davis Weather Station is a stand-alone system consisting of a tripod mounted data collection unit and a base receiver-display unit. The tripod mounted data collection unit houses the wind speed and direction, temperature, and humidity sensors. The base unit contains additional temperature and humidity sensors along with a barometric pressure unit. This base unit processes the information received from all the sensors and displays the output on a digital screen.

After installation and integration of the above equipment, Mobile Lab Unit No. 6 underwent field tests on roadways in the Santa Barbara area. All of the equipment was verified to be in working order and ready for the trip to Washington. As a final precaution against possible problems, extra supporting equipment and additional copies of all the software were included in this unit's equipment list.

4.0 SITE SELECTION

Site selection began early in the preparation stages of this study. The Washington Department of Transportation, city, and county traffic engineering officers were contacted to gather information on traffic patterns and ADT volumes for the three-county area. This preliminary data provided a starting point to narrow the scope of possible collection sites. The basic goal of the site selection phase of this study was focused on selecting sites to meet the criteria of the V*IDOE while keeping within the guidelines of the basic site selection criteria established by Hughes on previous studies.

The requirements of the WDOE study had two main goals:

I Yield 50,000 or more valid measurements from the "In Area" fleet. The "In Area" fleet was defined by the license plate database to be supplied by the WDOE. This database was comprised of vehicles registered within the state's non-attainment area that had an established history with the state's I/M contractor by previous visits, and remained liable for periodic emissions testing.

2. Yield 5,000 or more vehicles with three or more valid readings taken at different times.

The basic site selection criteria developed by Hughes can be categorized as follows:

1. The operator and equipment are safe throughout set-up, operation, and tear-down phases.

2. The operation does not present a safety hazard to the driving public. Minimal coning is required - the site is normally a single lane. This minimizes alteration of the motorist's driving pattern and enables an unobtrusive placement of the remote sensing equipment.

3. There are no nearby traffic devices (stop lights, stop signs, etc.) to alter driving conditions, i.e., a forced acceleration or deceleration, or traffic backing up into our monitoring site. However, there are occasions where a traffic control device can be useful if it regulates

traffic flow into the sensor site. The periodic gaps are useful for calibration and other operator tasks. These factors would be weighed site-by-site.

4. The road climbs gradually into the monitoring site in a sweeping bend. This tends to keep a driver maintaining a constant speed.

5. Natural and man-made material in close proximity to the monitoring equipment. Natural barriers and obstacles reduce the need for coning, offer better protection (trees, shrubs, utility poles, guardrails, etc.), and are less likely to cause a driver to change their normal driving behavior.

6. An "ideal site" would have a low likelihood of cold start conditions.

These criteria can be summarized as a safe, single lane, smooth traffic flow, slight upgrade with natural foliage (or barriers).

During the week prior to the start of data collection, John Brown and Guy Smith arrived in the Seattle area to continue the site selection process. Contact with the Washington Traffic Engineering Department proved to be the best contact for information on local traffic patterns and ADTs for the three-county area. We gained valuable information on hundreds of locations in the general area. Maps, ADT volume reports, automatic sensor locations, and a complete printout of all supporting data were made available for the three-county area. The amount of information available was so extensive, it required scheduled computer time to compile the information and output it in a useful format. This process took an additional week to be delivered. Unfortunately, this massive amount of information arrived too late to be of direct value to this study. However, this information is now available to the WDOE and Hughes, and will be a tremendous asset in selecting sites for future work in this area.

With the preliminary information received from the Department of Transportation and other area agencies, we began a three-day visual inspection of possible locations throughout the three-county area. At the end of this three-day effort, a list of 62 possible sites covering the three-county area was delivered to the WDOE for suggestions and approval. From this list of 6 possible sites, approximately 20 sites were determined to be the best candidates for collection. These final 20 sites met most of the criteria for both the VRDOE and Hughes requirements.

In order to achieve the goals of the WDOE, the decision was made to concentrate our efforts on the freeway on-ramps and interchanges in the three-county area. Only those sites with higher commute traffic ratings would be considered for use. During the collection phase of this study, some of the sites selected did not prove to be worth repeat visits and were dropped from the list.

Two reports have been prepared based on information extracted from the SLOOKUP.DBF and the WASHSITE.DBF databases. The first report, Washington Site List (Appendix C), contains information on the original list of 62 sites selected for this study. Along with the site location and site number, you will find additional information about each site. The second report, Sites Monitored (Appendix D), contains information on the sites used. This information includes dates monitored, collection times, operator, site number, and site location text.

5.0 PERMITTING

Permitting activities for this study began as the preliminary site location list began to take form. WDOE provided Hughes with a list of possible contacts in the agencies responsible for issuing permits in the three-county area. These agencies were contacted and all available information on site locations and collection time schedules were discussed. A point of contact was made with each of the responsible agencies and also with the Washington law enforcement agencies in the three-county area. When the final site list was available, each agency contact received a copy for review and suggestion. At that time, each agency determined what action would be needed to conduct our study in their respective areas.

WDOE also provided Hughes with a letter stating the work to be performed and identifying a contact name within WDOE. This letter was intended to serve as an introduction and brief explanation of our presence on the roadways in the three-county area.

The Washington Department of Transportation required no formal permit based on their understanding that our study did not involve construction or alteration of any roadway traffic lanes, and there would be only minimal impact on normal traffic flow. Therefore, the only requirement would be to follow safe roadside procedures and notify the proper agency if problems arose during collection. A "Site Locations and Times Schedule" was faxed to each of the agencies on a weekly basis. This notification was usually sent late in the week giving details for the following week's scheduled activities. Additional faxes were sent in the event of schedule changes. The following is a list of the contacts for each agency.

1. Kerry Swayne	Washington Dept. of Ecology
2. Roger Stienert	Washington Dept. of Transportation (North Area)
3. Frank Newboles	Washington Dept. of Transportation (Pierce Co.)
4. Lt. Hurlbut	Washington State Police, Traffic Dept.

Our site locations were visited by various representatives of the WDOE, Washington Department of Transportation, and Washington State Patrol during the course of this five-week collection period. The study concluded with no involvement by any agency. There were no traffic problems reported and no concerns expressed by law enforcement officers.

6.0 DATA COLLECTION

The data collection phase of this study began on 26 August, 1996. For our first effort, we selected a site expected to be a lower volume, less congested location. This selection would allow us to monitor the local driving habits and adjust the equipment setup procedure to create as little impact as possible on the normal traffic patterns. It also allowed us an opportunity to plan the best configuration of safety devices such as warning signs and cones. Previous studies have shown that the site setup configuration plays an important part in collecting data. For example, adjusting the position of safety cones on each side of the traffic lane has a direct affect on the speed and angle the average vehicle will take when approaching the site. Making small changes in the width of the lane can mean the difference between excessive braking and a smooth flow. This becomes extremely important when operating in heavy traffic areas where excessive braking can create a total stoppage of traffic.

The first two sites selected for this study yielded the expected results of lower volumes and moderate congestion. The equipment proved to be operating in a normal manner and adjustments to the setup procedure had been made to best suite the local driving habits. Data collection continued on schedule for the first two weeks. We then moved on to a location with much higher expected traffic volumes.

As data collection continued into the third week, preliminary data on the first two weeks of collection began to show an interesting trend. While the average traffic volumes were at an expected level, the number of "In Area" vehicles and repeat measurements were much lower than had been anticipated. Since WDOE required that only vehicles registered within the state's non-attainment area and liable for periodic emissions testing would be considered "valid," this trend was cause for concern. While Hughes is familiar with collection rates of our equipment in terms of the specific elements required, we *Were uncertain of the impact of these unique additional parameters. Considering the locations selected for emissions collection were believed to be well within the non-attainment areas, we were somewhat concerned at the volume of vehicles lost during collection as "invalid" based on registration records of vehicles previously tested at area UM facilities. Further data analysis over the course of the study revealed we were losing about 36% of the daily collection volume due to invalidation based on the "In Area" requirement. It became clear that in order to meet the goal of three separate measurements on 5,000 vehicles, our efforts in the final two weeks of collection would have to be concentrated on a single site with a history of high commute traffic.

The decision was made to return to a site we previously monitored that appeared to meet the requirements. Our efforts during the last week of this study were concentrated on site number 3, Highway 520 West bound onto Interstate 405 South. This proved to be a wise decision. The final data analysis showed we easily exceeded the 50,000 total valid vehicle requirement and were very near achieving the 5,000 vehicles with three or more detections.

Since there is a portion of vehicles registered in the non-attainment area that are not yet a part of the available I/M tested fleet (such as newer model vehicles not yet requiring testing), we also evaluated the multi-detection requirement on all Washington state vehicles with valid license plates. We termed this format to be "No-Bouncers." Using the No-Bouncer format, we do exceed the 5,000 vehicle, three or more detection requirement.

Data collection ended on 30 September, 1996. On 1 October, 1996, the unit departed the Seattle area for the return trip to Santa Barbara.

7.0 CLOSING COMMENTS AND OBSERVATIONS

The data collected for this study has been delivered to WDOE. Several attempts were required to provide a format that was complete and could be read by WDOE. There are a few points that should be mentioned based on our preliminary analysis of the data and observations from the actual data collection process.

In a letter, dated 21 October, 1996, John Brown of Hughes Technical Services Company informed WDOE of the completion of the data collection phase of this study and provided some preliminary statistics. Mr. Brown addressed the issue of the "invalid" data lost due to the high number of out-of-area registered vehicles. It was observed that approximately one out of every three vehicles passing through our sites on a daily basis were out-of-area vehicles. Mr. Brown suggested this may be an indication of the need to expand the area within which emissions testing is required. It is also apparent that the residents of the state are probably highly mobile with a significant volume of long distance commuters. One other possibility is that there is some registration fraud to avoid emissions testing. These observations are based on the fact that the daily percentage of invalid records was consistent over all of the sites monitored during this study. These are important factors and should be taken into consideration for future studies.

The 21 October letter also indicated that 16,000 vehicles were captured in a single collection day. Further review of the data revealed an error had been made and one day's collection was entered twice. The double entry also made it appear in the preliminary data that we would meet the 5,000 vehicle multi-detection requirement. The revised data set eliminates this redundant entry.

Table 7-1 may help to show the statistical variances between the "Bouncer or In Area" record and the "No-bouncer or complete Washington state license plate" record criteria.

One possible alternative to reduce the number of out-of-area vehicles may be by reducing the size of the geographic area covered by a collection site location. By using the registration database of addresses of the vehicle owners within the non-attainment area, it may be possible to target smaller areas and concentrate on locations of high traffic volumes to increase the daily valid collections.

Table 7-1. Comparison Table of Data Calculated from the "Valid" Records to "Invalid" Records *

	Using all Wash. as Valid Records	Using WDOE Criteria
TOTAL RECORDS COLLECTED	201,581	201,581
AVERAGE COLLECTION PER DAY	7,753	7,753
HIGHEST DAY COLLECTION	13,185	13,185
LOWEST DAY COLLECTION	2,233	2,233
TOTAL INVALID CO	22,013	22,013
AVERAGE LOSS PER DAY COLLECTION	846	846
AVERAGE DAILY LOSS PERCENTAGE	10.91%	10.91 %
TOTAL INVALID SPEED	5,840	5,840
AVERAGE LOSS PER DAY	225	225
AVERAGE DAILY LOSS PERCENTAGE	2.89%	2.89%
TOTAL VALID RECORDS	137,523	64,028
AVERAGE PER DAY	5,289	2,463
HIGHEST DAY	9,561	4,406
LOWEST DAY	1,602	731
AVERAGE DAILY PERCENTAGE	67.29%	31.38%
TOTAL COST		
PER RECORD – **ALL**	$0.24	$0.24
PER **VALID** RECORD	**$0.35**	**$0.76**

* Note: “Valid” records data uses the criteria of "in Area or Valid License Plate" established by WDOE
"Invalid" records data uses the criteria of all Washington state license plates considered as valid

Working with the Washington State Department of Ecology and all the other organizations involved with this study has been a very pleasant experience. We at Hughes would like to extend our sincere thanks to all who have contributed in bringing this project to a successful conclusion. You may contact those of us involved in the program for follow-up questions at the following addresses:

Guy Smith, HTSC Van Operator
Santa Barbara Research Center
Bldg. B6, MS 84
75 Coromar Drive
Goleta, CA
Phone (805) 562-2365
Page (800) 560-6412
Fax (805) 562-4518
e-mail: gasmith1@ccgate.hac.com

John Brown, HTSC Program Manager
Hughes Technical Services Company
1201 W. N. Carrier Parkway, STE 100
Grand Prairie, TX 75050
Phone (972) 988-4013
Pager (800) 983-2511
Fax (972) 988-4022
e-mail hjbrown@ccgate.hac.com

Susan Reed, HTSC Systems Administrator
Hughes Technical Services Company
3815 E. Grove St, STE 3
Phoenix, AZ 85040
Phone (602) 470-1986
Pager (602) 599-6576
Fax (602) 470-1935
e-mail sereed@ccgate.hac.com

Michael Gray, SBRC Program Manager
Santa Barbara Research Center
Bldg. B6, MS 84
75 Coromar Drive
Goleta, CA
Phone (805) 562-2358
Fax (805) 562-4518

Sylvia Salazar, SBRC Contracts
Santa Barbara Research Center
Bldg. B6, MS 11
75 Coromar Drive
Goleta, CA
Phone (805) 562-4398
Fax (805) 562-4518

Thomas (Lindy) Heidt, SBRC Marketing
Santa Barbara Research Center
Bldg. B6, MS 84
75 Coromar Drive
Goleta, CA
Phone (805) 562-4724
Fax (805) 562-4518

Jay Peterson, SBRC Systems Engineer
Santa Barbara Research Center
Bldg. B6, MS 84
75 Coromar Drive
Goleta, CA
Phone (805) 562-2439
Fax (805) 562-4518

APPENDIX A

STATS.DBF

COLLECTION STATISTICS REPORT

Date	Site No	Time	Temp	Humidity	Wind Spd	Wind Dir	Cond
08/26/1996	56	08:13 AM	67	50	0		CLEAR
08/26/1996	56	12:00 PM	82	34	0		CLEAR
08/26/1996	56	02:47 PM	82	34	0		CLEAR
08/27/1996	56	06:50 AM	60	74	0		CLEAR
08/27/1996	56	11:32 AM	67	60	4	E	CLEAR
08/28/1996	56	07:00 AM	61	74	0		CLEAR
08/28/1996	56	01:54 PM	83	35	0		CLEAR
08/29/1996	57	08:20 AM	75	47	2	NW	CLEAR
08/29/1996	57	10:44 AM	80	39	0		CLEAR
08/30/1996	57	07:55 AM	64	82	0		OVERCAST
08/30/1996	57	11:08 AM	81	42	I	SE	CLEAR
09/02/1996	57	08:26 AM	54	75	0		CLEAR
09/02/1996	57	12:30 PM	72	39	0		CLEAR
09/02/1996	57	04:33 PM	79	34	0		CLEAR
09/03/1996	61	06:27 AM	61	64	0		OVERCAST
09/03/1996	61	09:30 AM					RAIN
09/04/1996	61	07:04 AM	52	85	0		FOG
09/04/1996	61	09:30 AM	58	76	0		OVERCAST
09/05/1996	61	06:18 AM	55	78	0		OVERCAST
09/05/1996	61	10:30 AM	62	65	I	SE	CLEARING
09/06/1996	61	06:22 AM	56	73	0		OVERCAST
09/06/1996	61	10:00 AM	69	48	0		CLEAR
09/09/1996	61	06:10AM	54	71	0		CLEAR
09/09/1996	61	09:48 AM	69	49	0		CLEAR
09/10/1996	3	06:30 AM	56	75	0		CLEAR
09/11/1996	3	06:50 AM	58	73	0		CLEAR
09/11/1996	3	0 1:00 PM	74	51	0		SUNNY

09/12/1996	3	07:00 AM	59	70	0		CLEAR
09/12/1996	3	09:30 AM	62	61	0		SUNNY
09/13/1996	3	06:44 AM	59	68	0		OVERCAST
09/16/1996	24	07:30 AM	48	78	1	SW	SUNNY
09/16/1996	24	09:30 AM	56	60	0		SUNNY
09/16/1996	24	02:00 PM	62	45	6	SW	CLEAR
09/17/1996	24	07:15 AM	55	77	1	SW	OVERCAST
09/17/1996	24	09:07 AM	57	73	0		OVERCAST
09/18/1996	8	08:33 AM	49	56	0		OVERCAST
09/18/1996	8	01:00 PM	63	61	0		CLEAR
09/18/1996	8	02:30 PM	58	78	0		OVERCAST
09/18/1996	8	03:00 PM	62	70	0		RAIN
09/19/1996	8	06:42 AM	50	'80	0		OVERCAST
09/19/1996	8	11:40 AM					RAIN
09/19/1996	8	02:30 PM	61	78	0		CLEARING
09/20/1996	8	07:21 AM	49	75	0		OVERCAST
09/20/1996	8	01:30 PM	61	78	0		RAIN
09/23/1996	3	06:55 AM	47	71	0		CLEAR
09/23/1996	3	09:47 AM	58	50	2	N	CLEAR
09/24/1996	3	06:45 AM	56	63	0		CLEAR
09/25/1996	3	07:14 AM	55	62	0		CLEAR
09/25/1996	3	12:21 PM	67	45	1		CLEAR
09/26/1996	3	07:04 AM	51	74	0		CLEAR
09/26/1996	3	09:57 AM	67	46	0		CLEAR
09/27/1996	3	06:53 AM	50	81	0		CLEAR
09/30/1996	3	08:40 AM	53	78	0		OVERCAST
09/30/1996	3	01:30PM	68	57	1		CLEAR

APPENDIX B

IHTCOUNT.DBF

VALID RECORD RUNNING TOTAL REPORT

VALID RECORDS BREAKDOWN

(USING D.O.E. SUPPLIED LICENSE PLATE DATABASE)

DATE	RUNNING TOTAL	1X	2X	3X	4X	5X	6X	7X	8X	9X	10X	11X	12X	13X	*14X*	15X
08/26/1996	1052	1032	11	1	0	0	0	0	0	0	0	0	0	0	0	0
08/27/1996	2418	1935	213	13	2	2	0	0	0	0	0	0	0	0	0	0
08/28/1996	3848	2549	464	104	7	5	1	0	0	0	0	0	0	0	0	0
08/29/1996	5455	4013	522	113	7	5	1	0	0	0	0	0	0	0	0	0
08/30/1996	6505	4760	643	133	6	6	1	0	0	0	0	0	0	0	0	0
09/02/1996	8249	6100	794	154	13	7	2	0	0	0	0	0	0	0	0	0
09/03/1996	9075	6924	795	154	13	7	2	0	0	0	0	0	0	0	0	0
09/04/1996	11101	8423	1054	157	13	7	2	0	0	0	0	0	0	0	0	0
09/05/1996	13508	9711	1431	267	18	10	2	0	0	0	0	0	0	0	0	0
09/06/1996	17022	11159	1917	508	96	19	2	2	0	0	0	0	0	0	0	0
09/09/1996	20351	12192	2238	763	219	82	14	1	1	1	0	0	0	0	0	0
09/10/1996	23904	14578	2487	822	290	103	26	4	1	1	1	0	0	0	0	0
09/11/1996	27406	15976	3071	940	320	148	51	14	2	2	1	0	0	0	0	0

DATE	RUNNING TOTAL	1X	2X	3X	4X	5X	6X	7X	8X	9X	10X	11X	12X	13X	*14X*	15X
09/12/1996	30834	17078	3391	1251	378	172	82	34	9	3	2	0	0	0	0	0
09/13/1996	31565	17122	3393	1283	455	186	92	41	19	4	2	1	0	0	0	0
09/16/1996	33559	19013	3443	1284	455	186	92	41	19	4	2	1	0	0	0	0
09/17/1996	35426	20164	3753	1312	458	186	92	41	19	4	2	1	0	0	0	0
09/18/1996	38976	23486	3856	1318	459	186	92	41	19	4	2	1	0	0	0	0
09/19/1996	41393	24640	4430	1348	464	187	92	41	19	4	2	1	0	0	0	0
09/20/1996	43971	25701	4841	1557	477	189	93	41	19	4	2	1	0	0	0	0
09/23/1996	48372	27419	5184	1686	626	282	113	63	38	15	3	0	1	1	0	0
09/24/1996	51743	28349	5504	1748	705	378	169	79	52	30	14	1	0	0	2	0
09/25/1996	55140	29231	5758	1887	749	418	238	129	74	34	23	12	2	0	0	2
09/26/1996	58792	30114	6066	2004	796	458	284	180	121	53	24	21	10	2	0	2
09/27/1996	61742	30860	620	206	867	494	312	200	155	88	47	16	18	8	3	2

			6	1												
09/30/1996	63974	31419	6366	2166	916	542	315	207	174	106	50	25	24	4	7	3

VALID RECORD REPORT PAGE 2

APPENDIX C

SLOOKUP.DBF

WASHINGTON SITE LIST

Site No.	Location	Comments
1	124 TH AVE TO 520 WEST	UP GRADE RT TURN
2	520 WEST TO 1 405 NORTH	HEAVY TRAFFIC
3	520 WEST ONTO 1 405 SOUTH	WIDE LANE USE AREA CLOSE TO EXIT POINT
4	90 EAST TO I 405 SOUTH	
5	1405 SOUTH TO I 5 SOUTH	INTERCHANGE, HEAVY TRAFFIC
6	518 EAST TO 1 5 SOUTH	LONG UP HILL INTERCHANGE
7	15 SOUTH TO 512 EAST	CRUISE
8	512 WEST TO 1 5 NORTH	ACCEL
9	512 WEST TO 1 5 SOUTH	CRUISE
10	16 WEST TO SOUTH 19TH ST	OFF RAMP CRUISE
11	16 WEST TO 6TH AVE	OFF RAMP
12	PEARL ST ONTO 16 WEST	ONLY IF ADT IS HIGH
13	JACKSON AVE OFF FROM 16 WEST	LONG SWEEPING BLIND RAW, ADT MUST BE HIGH TO DO
14	JACKSON AVE OFF FROM 16 EAST	UP HILL GRADE, DO ONLY IF TRAFFIC ADT IS HIGH
15	JACKSON AVE ONTO 16 EAST	DOWN HILL RAW, ONLY IF ADT IS HIGH
16	UNION ONTO 16 EAST	VAN CLOSE TO FREEWAY ENTER POINT, ONLY IF ADT IS HIGH
17	16 EAST OFF TO SPRAGUE	FLAT, CRUISE 40
18	16 EAST OFF TO 38TH ST	CRUISE 35
19	I 5 SOUTH OFF TO 512 EAST	ACCEL 45 INTERCHANGE
20	15 NORTH TO 512 EAST	CRUISE 40
21	512 EAST OFF TO 9 TH ST SW	ACCEL35
22	9 TH ST SW ONTO 512 WEST	ACCEL35
23	161 TO 512 EAST(NORTH)	LT UP FULL 45 ACCEL
24	161 NORTH TO 167 NORTH	FLAT CRUISE 45, INTERCHANGE
25	167 SOUTH TO 4 1 0 EAST	CRUISE 45
26	161 NORTH TO 167 SOUTH	LT UP HILL ACCEL 45
27	167 NORTH TO 512 WEST	FLAT CRUISE 45
28	167 NORTH TO 18 EAST	LONG RAW. CRUISE 45
29	15 TH ST ONTO 167 NORTH	CRUISE 45
30	18 EAST ONTO 167 NORTH	ACCEL40
31	18 WEST TO 167 NORTH	LONG SWEEPING TURN TO STRIGHT AT END ACCEL 40
32	167 SOUTH TO 18 WEST	CRUISE 45, HIGH TRAFFIC AREA
33	18 EAST TO 167 NORTH	
34	NERIDIAN ONTO 512 (161) NORT14	LT UP ACCEL 35
35	15 NORTH ONTO 18 EAST, EXIT # 142A	CRUISE 45
36	18 EAST ONTO 1 5 NORTH	45 ACCEL
37	18 WEST TO 1 5 NORTH	40 ACCEL, LT UP

Site No.	Location	Comments
38	272 ND ST ONTO 1 5 NORTH	45 ACCEL
39	516 TO 15 NORTH	35 ACCEL IF ADT IS f IIGH
40	SOUTH 188 TH ST ONTO 1 5 NORTH	35 ACCEL IF ADT IS HIGH
41	1405 SOUTH TO 15 NORTH	SHORT DOWN DECEL, NOT GOOD
42	1 5 SOUTH OFF TO PACIFIC HWY S,EXITI 58	MOSTLY COAST MODE
43	15 NORTH OFF TO COLUMBIA, EXIT 163	LT UP, 40 ACCEL, DO IF ADT IS HIGH
44	1 5 NORTH TO 90 EAST, EXIT 164	ACCEL 45 LY UP, HALF WAY UP RAMP, TRAFFIC COULD BACKUP
45	I 5 NORTH TO 522 EAST	SLOW SPEED CURVE FLAT, DO AT FAR END WIDE AREA
46	244 TH ST ONTO 1 5 NORTH	FLAT SHORT RAW ACCEL 35, IF ADT IS HIGH
47	1 5 NORTH TO 1 405 SOUTH, EXIT 182	45 CRUISE GOOD AM, DO AT FAR END
48	527 TO 1405 NORTH,EXIT 26 ON 405	CRUISE 40
49	1405 NORTH TO 527	45 CRUISE, FLAT
50	1405 SOUTH TO I 5 SOUTH	CRUISE 50 DO AT FAR END CLOSE TO 1 5
51	15 NORTH TO 525 NORTH(WEST)	45 CRUISE, DO AT CONSTRUCTION BARRELS
52	164 TH ST ONTO 1 5 NORTH,EXIT 183	DO AT FAR END,45 CRUISE, LT UP
53	527 WEST ONTO 1 5 NORTH	DO AT FAR END CRUISE 45
54	2 EAST TO 204 SOUTH	DO NOT DO, OUT OF AREA TRAFFIC PER DOE, FLAT 45
55	2 WEST ONTO I 5 SOUTH	DO NOT DO, OUT OF AREA TRAFFIC,PER DOE ACCEL 35
56	15 SOUTH OFF TO 526 WEST, EXIT 189	HIGH AM TRAFFIC, BOENG, UP HILL 45 ACCEL
57	1 5 SOUTH OFF TO 99(527) EXIT 189	LT UP, HIGH TRAFFIC TO MALL
58	527 TO 1 5 NORTH	ONLY IF ADT IS HIGH, DOWN 35 DECEL POSSSIBLE
59	DOWNTOWN (SPOKANE ST) ONTO 1 5 NORTH	SURFACE ST ONTO I 5 COMES FROM UNDER 99 EAST AT EX 163
60	DOWNTOWN (SPOKANE ST)ONTO 1 5 SOUTH	SURFACE ST TO 1 5 UNDER 99 EAST EX 163, BACKUP LIKELY
61	520 WESTBOUND AT REDMONDS	DO AT CONSTRUCTION WALL OPENING
62	1405 SOUTH TO 520 WEST	HIGH TRAFFIC, BACKUP LIKELY AT PEAK HRS

APPENDIX D

WASHSITE.DBF

WASHINGTON STUDY SITE LIST REPORT SITES MONITORED

Date	Location	**Start-time**	**End-time**	Operator	**Van**	Tx-loc
08/26/1996	56	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 526 WEST, EXIT 189
08/27/1996	56	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 526 WEST, EXIT 189
08/28/1996	56	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 526 WEST, EXIT 189
08/29/1996	57	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 99(527) EXIT 189
08/30/1996	57	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 99(527) EXIT 189
09/0211996	57	05:00:00	19:00:00	Smith	6	15 SOUTH OFF TO 99(527) EXIT 189
09/03/1996	61	05:00:00	19:00:00	Smith	6	520 WESTBOUND AT REDMONDS
09/04/1996	61	05:00:00	19:00:00	Smith	6	520 WESTBOUND AT REDMONDS
09/05/1996	61	05:00:00	19:00:00	Smith	6	520 WESTBOUND AT REDMONDS
09/06/1996	61	05:00:00	19:00:00	Smith	6	520 WESTBOUND AT REDMONDS
09/09/1996	61	05:00:00	19:00:00	Smith	6	520 WESTBOUND AT REDMONDS
09/10/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO I 405 SOUTH
09/11/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1405 SOUTH
09/12/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1 405 SOUTH
09/13/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1405 SOUTH
09/16/1996	24	05:00:00	19:00:00	Smith	6	161 NORTH TO 167 NORTH
09/17/1996	24	05:00:00	19:00:00	Smith	6	161 NORTH TO 167 NORTH
09/18/1996	8	05:00:00	19:00:00	Smith	6	512 WEST TO 1 5 NORTH
09/19/1996	8	05:00:00	19:00:00	Smith	6	512 WEST TO 1 5 NORTH
09/20/1996	8	05:00:00	19:00:00	Smith	6	512 WEST TO 1 5 NORTH
09/23/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1 405 SOUTH
09/24/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1405 SOUTH
09/25/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO I 405 SOUTH
09/26/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1 405 SOUTH
09/27/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO 1 405 SOUTH
09/30/1996	3	05:00:00	19:00:00	Smith	6	520 WEST ONTO I 405 SOUTH

Remote Sensing Data Analysis Report

Jacob A. Wegelin
Paul D. Sampson
Center for Statistical Consulting
University of Washington
Box 354322
Seattle WA 98195

February 2, 1998

Abstract

Our goal was to determine whether measurements of CO and HC emissions obtained by remote sensing, could be used to determine accurately the emission levels of individual vehicles. In particular we sought to determine whether remote sensing data could be used to Identify gross emitting vehicles and clean vehicles. We took as our gold standard the Emission Check data obtained by the Washington Department of Ecology as part of the legally mandated Emission Check program. We created a subset (5607 measurements on 1301 vehicles) of a remote sensing dataset acquired by the Santa Barbara Research Center of Hughes Aircraft Company [9] and compared it with Emission Check (EC. Correlations between remote sensing, data (RSD) and EC are positive but small. Prediction intervals, even 50% prediction intervals, encompass nearly the entire range of EC values and hence are useless. This is due in large part to the fact that the within-vehicle variation of RSD is high. After appropriate transformation and after correction for differences in acceleration repeated remote sensing measurements taken on the same vehicle have an interquartile range that is over 70% of the interquartile range of the data as a whole. Furthermore between-vehicle variation is not negligible. Even when repeated remote sensing measurements oil individual vehicles have been adjusted for acceleration and summarized to yield a single more reliable estimate for each vehicle correlations between these RSD summaries and EC remain low. They range from 0.30 for vehicles on which 8 repeated remote sensing measurements were obtained. Using conventional methods of inverse regression, it is impossible to predict the EC value of a vehicle from RSD in any meaningful way.

Chapter 1

Summary of the Analysis

Prologue **Much** of this report deals with activities that could be called overhead-reading in data, exploring and describing data, cleaning data, finding all appropriate transformation for the data, choosing all appropriate subset of the data on which to work, merging data. We discuss these tasks in the Appendices (A, B, C, D, and E). In the numbered chapters we discuss only the actual analysis. The heart of the analysis was a comparison of RSD and EC, and specifically an attempt to predict EC from RSD. An additional task falling under the heading of "Analysis," however, was the computation of within-vehicle summaries of RSD. We did this to take advantage of the fact that we had multiple RSD hits for many vehicles. By computing summaries we sought to reduce the variation of the RSD values, and thereby to increase our chances of getting a close correlation between them and EC.

1.1 Results

It is easy to see from a couple scatterplots that there is too much variation In RSD for us to be able to predict EC from RSD in any meaningful or useful way. What we see in scatterplots, we can confirm more formally by going through the computations involved in inverse regression. When we say that we cannot predict EC from RSD in any meaningful or useful way, we are referring to the size of the prediction Interval, the degree of uncertainty in our prediction. The uncertainty is so great that prediction intervals in many cases encompass the entire range of the data. This is true even when we reduce the variance of RSD by computing a single summary value for each vehicle. We give a "nutshell" discussion of this in Chapter 1.2. We approach the subject more formally in Chapter 5.

1.2 A Nutshell Discussion of Our Results

Please consider Figure 1.1 and Figure 1.2. These show the data from all vehicles for which

- the Emission Check date occurred within 90 days of at least one remote sensing date.

- we had 3 or more remote sensing hits, during the 90-day period. (We call RSD hits that occurred during the 90-day period RSD90HITs.)

Emission check values are plotted on the x-axis, RSD summaries on the y-axis. (We discuss RSD Summaries in Chapter 2. They can be thought of as means computed separately for each vehicle from all the RSD90HITs for that vehicle, adjusted for the fact that each RSD90HIT had a different acceleration.) The original scale is displayed on the bottom and left sides, the log-transformed scale on the top and right. The correlation between log-transformed EC and log-transformed RSD is noted on the plot.

The solid, slanted line represents the best linear fit between emission check and adjusted, summary RSD. This is known as the linear regression line. It represents our best attempt to explain remote sensing data in terms of Emission Check. When we say that there is too much variation in RSD, we refer to the fact that the vertical scatter around the linear regression line is too great.

Vertical scatter at any value on the x-axis (at any given value for EQ represents the uncertainty in RSD summaries for vehicles of that given EC value. For us to be able to predict EC from RSD, we need the vertical scatter around the linear regression line to be tight, relative to the steepness of the line. The steeper the line, the less tight the vertical scatter would need to be. Tight scatter relative to the steepness of' tile line corresponds to correlation close to 1 or -1. In our case, correlations are 0.36 for both CO and HC. This relatively small correlation corresponds to a wide scatter around, a fairly flat regression line.

Large vertical spread (wide vertical scatter) around tile regression line translates into large uncertainty in tile EC value that we might hope to predict from an RSD summary. Large uncertainty in the predicted EC value means a wide prediction interval. For any given RSD value, as the vertical spread gets bigger, at some point the prediction actually breaks down. At this point, the prediction interval begins to encompass the entire range of the data. Such an Interval is useless, of course.

We discuss tile prediction of EC from RSD in more detail in Chapter .5. In particular, we give a geometric explanation of how tile prediction breaks down.

1.3 Is the Fault All in RSD, or Also in EC?

Once we've determined that there is too much variation in RSD for us to be able to predict EC from RSD in any meaningful or useful way, the question remains

(Text continued on p. A-37)

Figure 1.1: Remote sensing CO summary as a function of Emission Check. . Data were log-transformed before plotting. Original scale can be read on the bottom and left, transformed scale on top and right. Only those vehicles are included for which the Emission Check date is within 90 days of the remote sensing date, and for which we have at least 3 remote sensing observations (RSD hits). Solid line is linear fit to the data, weighted by the number of remote sensing observations for each vehicle.

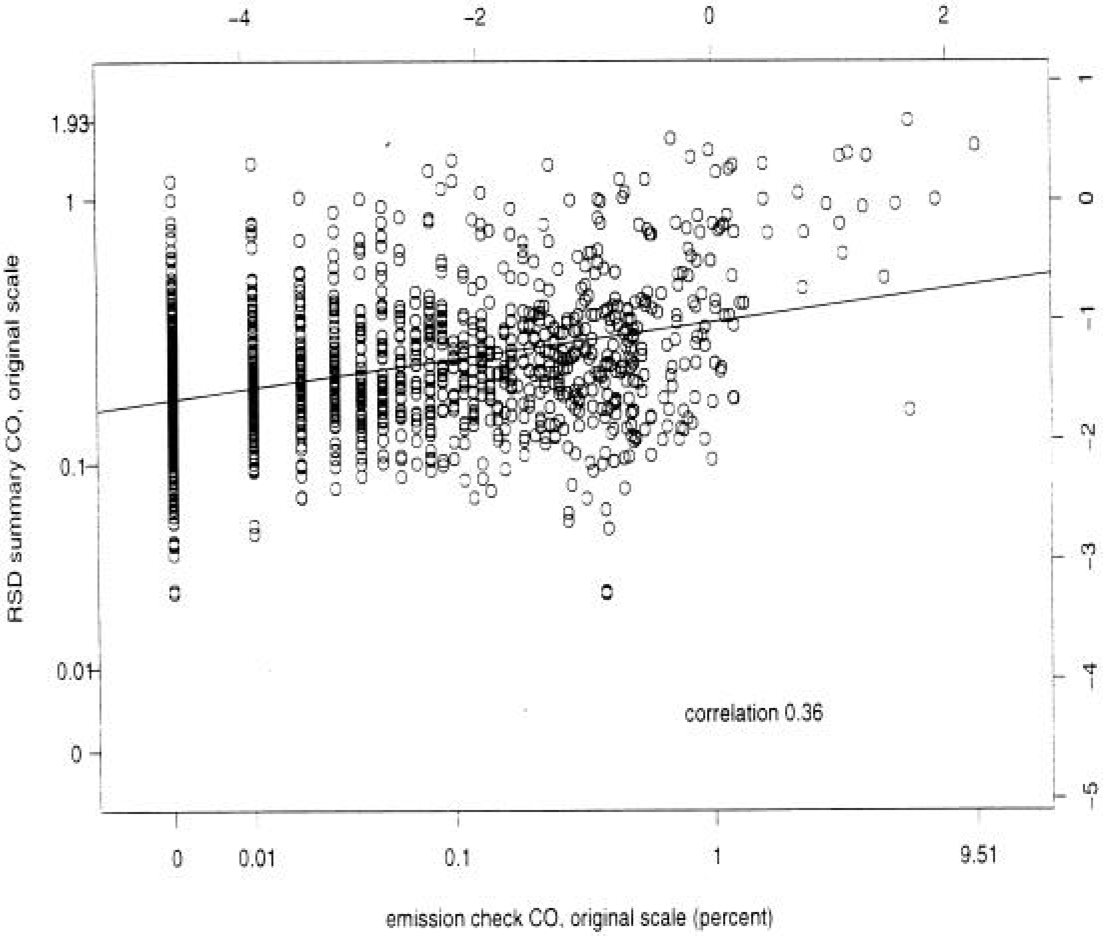
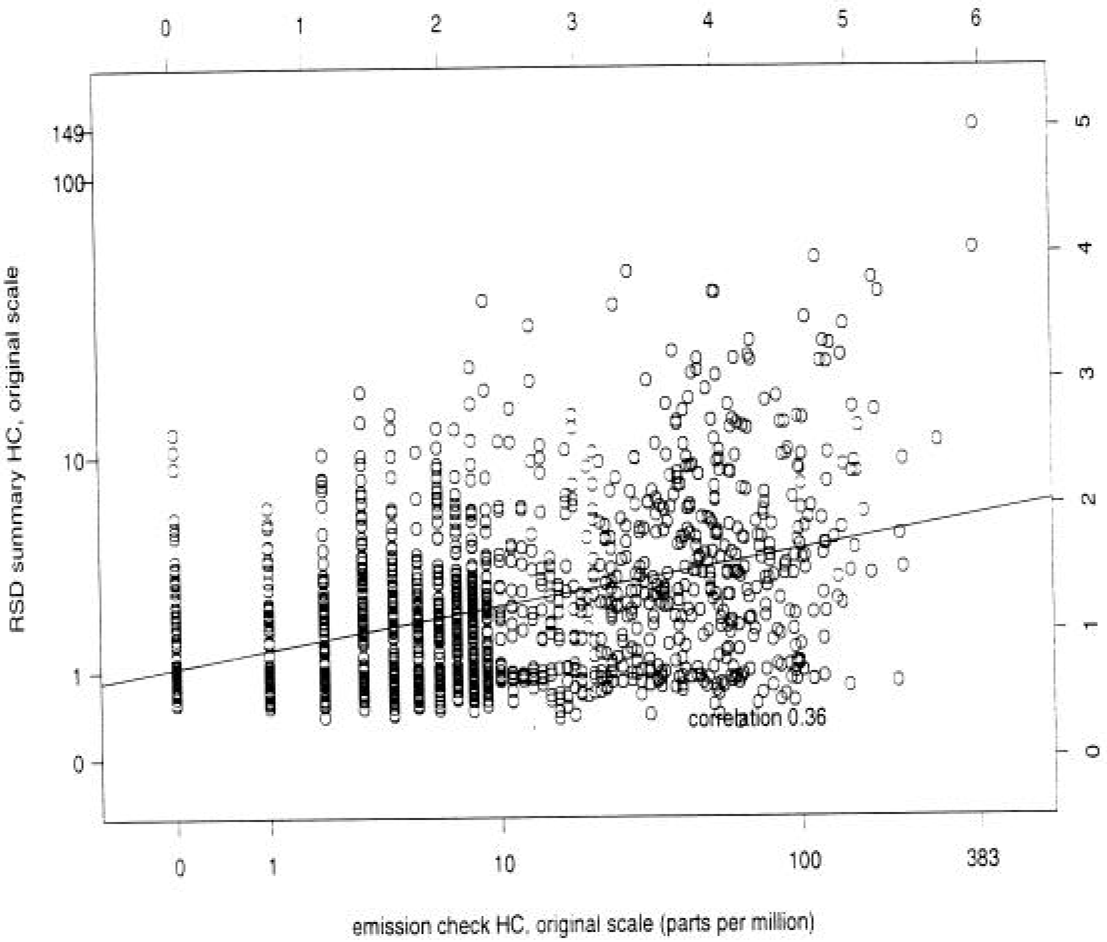


Figure 1.2: Remote sensing HC summary as a function of Emission Check. Data were log-transformed before plotting. Original scale can be read on the bottom and left, transformed scale on top and right. Only those vehicles are included for which the Emission Check date is within 90 days of the remote sensing date, and for which we have at least 3 remote sensing observations (RSD hits). Solid line is linear fit to the data, weighted by the number of remote sensing observations for each vehicle.



to what degree the fault lies in unreliable RSD values and to what degree it lies elsewhere. Although RSD values are not very reliable, as detailed in Chapter 2, our data suggest that a significant amount of the fault lies elsewhere. Our data suggest that, even if we had 1000 RSD hits on each vehicle, all taken within 90 days of the EC date, the scatter of the resulting RSD summaries around the linear regression line in Figures 1.1 and Figure 1.2, would still be too big for us to predict EC precisely. We should not be surprised at this because we have no reason to believe that EC itself is measured without error. Furthermore, we have no direct way in the data currently available to us (Detailed in Chapter A) to determine how large the variance is of EC measurements. From the very beginning it was our intention to take EC as a standard" (see our 9 May 1997 proposal.), but in fact EC may not be precise enough to deserve such a name. We return to this issue in chapter 1.4.

1.4 Within- Versus Between-Vehicle Variation

The distinction made in Chapter 1.3 between whether “the fault lies all in RSD" or “the fault lies elsewhere” is related to tile distinction between within-vehicle variation and between-vehicle variation. This pertains to a question raised implicitly by Kerry Swayne In his email of Thursday, 8 January 1998. The phrase “between-vehicle variation” refers to a feature of the data that is very much of interest to us.

We should be clear about the difference between “within-vehicle variation” and “between-vehicle variation."

- The phrase “within-vehicle variation” encompasses only the reliability of RSD. Reliability is also called repeatability (for instance, in Kerry Swayne's email). The inverse of reliability is variance or spread. The question, “How much within-vehicle variation is there is in RSD?” is equivalent to the question: “If we measure RSD on a single vehicle many times, how different will the measurements be from each other? Will they be spread out or tightly clustered?”

- The phrase “between-vehicle variation” includes uncertainty in our ability to correlate RSD with EC. The question, “How much between-vehicle variation is there in RSD?" is related to the following issue. Suppose we could take a very large number of RSD measurements within a short period of time, under almost constant conditions, on a sample of vehicles that all happen to have identical EC values. Suppose we get the same number of RSD hits on each vehicle. Suppose we then compute the average RSD for each vehicle, or the acceleration-adjusted RSD summary. (Acceleration adjusted RSD summaries are discussed in Chapter 4). How close to each other would the vehicle averages be? (Or, alternatively, how close to each other would the acceleration-adjusted RSD summaries be?)

If the only limiting factor in our ability to predict EC from RSD were the within-vehicle variance of RSD, then we could predict EC accurately and precisely from RSD in the hypothetical scenario described in the previous paragraph. The larger the number of hits on each vehicle, the closer the RSD summaries would be to each other. If we got enough hits, we could get Summaries so close to each other that they would be indistinguishable. As we shall see in Chapter 3, this is not the case.

1.4.1 The "True" RSD Value of a Vehicle

In Chapter 1.4 we introduced the notion of taking the average, or summary, of a very, large number of RSD measurements on the same vehicle. We think of this hypothetical number as the true RSD value for a vehicle. It is hypothetical because we call never get such a large number of measurements on the same vehicle. When we say within-vehicle variation is nonzero, we mean that several vehicles with identical EC values will have different true RSD values.

Chapter 2

Within-Vehicle Variation

When we examine within-vehicle variation of RSD, we are seeking to answer the following question: If we take several RSD measurements on the same vehicle and adjust for differences in conditions under which the measurements were taken, how similar will these measurements be to each other?

2.1 Residuals from the Linear Mixed-Effects Model

To examine the reliability of RSD, we took at **residuals.** Residuals are numbers that are left over after summary values have been subtracted from them. The residuals that concern us for purposes of examining RSD reliability are the log-transformed RSD values, after correction for acceleration and subtraction of the RSD summaries. We will call these "the residuals from the linear mixed-effects model," or simply "within-vehicle residuals," to distinguish them from "the residuals from regressing, RSD summary on EC," which we discuss in Chapter 3. You can see histograms of the residuals from the linear mixed-effects model in Figure 2.2, and in Figure 2.2.

By comparing the spread of the residuals from the linear mixed-effects model with the spread of the log-transformed data, as a whole, we can get an idea how reliable RSD are.

2.2 Details of How the Residuals from the Linear Mixed-Effects Model Were Computed

The reader might want to skip this on a first reading.

To get a visual idea of how the residuals from the linear mixed-effects model are computed, it might be useful to consider the process for one vehicle. We display the log-transformed RSD CO data for one vehicle in Figure 4.1.

Figure 2.1: CO residuals from the linear mixed-effects model.

Figure 2.1: CO residuals from the linear mixed-effects model.

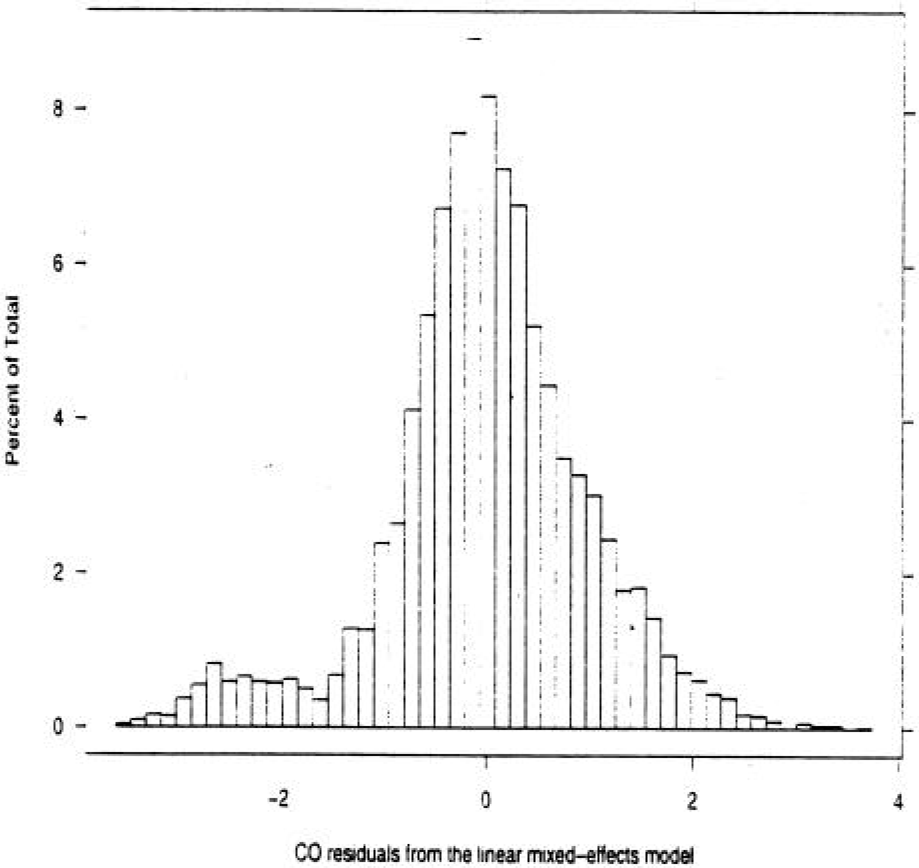
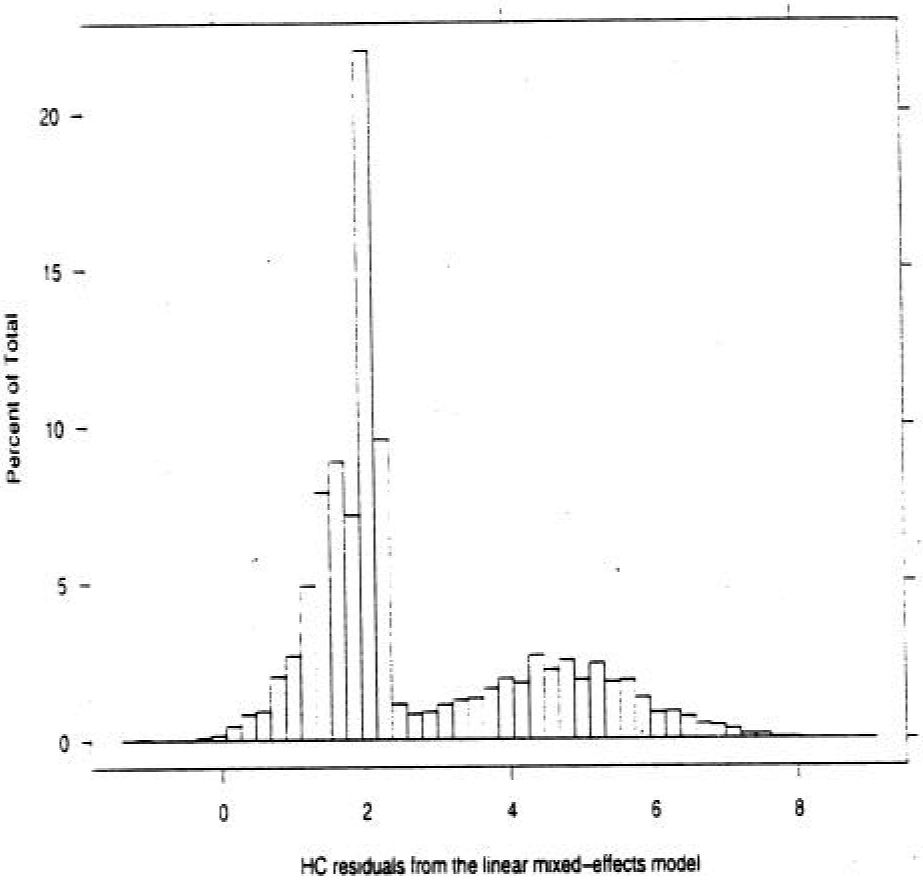


Figure 2.2: HC residuals, from the linear mixed-effects model. Because raw HC is more skewed than raw CO, this histogram deviates further from the bell-shaped (normal Gaussian) distribution than the corresponding histogram for CO. (We discuss the further in Chapters B.6.2 and C3.)



(We discuss the plot In Chapter 4.) For any of the 7 RSD90HITs displayed in this figure, the residual from the linear mixed-effects model is the difference between the following two quantities:

- The log-transformed RSD value.
- The height of the slanted solid line at the centered DELTA value for this particular log-transformed RSD value.

If you draw a vertical line segment from one of the plotting characters depicting a log-transformed RSD CO value to the slanted solid line, the length of this line segment is the value of the residual from tile linear mixed-effects model for this particular RSD hit.

2.3 Variation in CO Residuals

Twenty-five percent of the CO residuals from the linear mixed-effects model are below –0.47. This is the 25th percentile or the 0.25 *quantile*, or the first *quartile*. Twenty-five percent are above 0.544. This is the 75th percentile or the 0.75 quantile, or the third quartile. The median (the 50% point) is -0.0122. We call the difference between the first and third quartiles the interquartile range**.** Interquartile range is a convenient way to measure the spread or uncertainty in a set of data. By definition, the middle 50% of the data lie within the interquartile range. In this case the interquartile range for the residuals from the linear mixed-effects, model is 1.01. We can express this as all uncertainty of roughly

$$\pm 0.507. \qquad (2.1)$$

Recall that we are dealing with log-transformed RSD values at this point, so that the units are not enlightening. (The transformation, and its rationale, are discussed in Section B.3.) Only after we transform back into the original scale can we interpret this number.

We will do that now. This will give us a rough idea of what this spread means in practical terms. Suppose we measure a vehicle that has a true RSD emission CO of 0.21% on the original scale, which is the median for remote sensing CO. Note that we cannot obtain the true value by measurement. Each measurement contains a degree of uncertainty. The true value may be thought of as the average of a very large number of measurements of the same variable taken under identical conditions. We introduced this idea in Chapter 1.4.

This translates into

$$\log(0.01 + 0.21) = -1.51 \qquad (2.2)$$

on the transformed scale. By Our rough uncertainty estimate (2.1) we can expect that 50% of the time the measurement after log-transformation would be between -2.02 and -1.01.

Finally, we transform these values back to the original scale and find that 50% of the time the interquartile range of raw RSD CO measurements on this vehicle would be between

$$\exp(-2.02) – 0.01 = 0.123\% \qquad (2.3)$$

and

$$\exp(-1.01) - 0.01 = 0.355\%. \qquad (2. 1)$$

This means that 25% of the raw CO values for this vehicle will lie below 0.123% and 25% will lie above 0.355%.

We may compare the spread in the residuals from the mixed-effects model with the spread in the data as a whole. Because of the skewed nature of the raw data, this comparison must be made on the log-transformed scale, as with all our analyses. (For a discussion of skew and the need for the log transformation. See Chapter B.3.) The log-transformed remote sensing CO values have a median of -1.51. Twenty-five percent are below -2.04, twenty-five percent above -0.755. Thus their interquartile range is 1.29, which is about 1.27 times the interquartile range of the residuals from the linear mixed-effects model. Put differently, the interquartile range of the within-vehicle residuals is 78.9% of the interquartile range of the data as a whole.

2.4 Variation in HC Residuals

The histogram of the HC residuals is more skewed than the histogram of the CO residuals, as can be seen by comparing Figure 2.1 with Figure 2.2. Going through the same process as we went through In Chapter 2.3, we find that the quartiles of the HC within-vehicle residuals are

- 1.6.5 (first quartile 25%)
- 2.06 (second quartile median)
- 3.76 (third quartile 75%)

Thus the interquartile range for HC within-vehicle residuals 2.11 and by dividing this in half we get an uncertainty of roughly

$$\pm 1 .06 \qquad (2.5)$$

Again, recall that we are dealing with log-transformed RSD values so that the units are not enlightening.

Now the skewed nature of HC becomes evident, for the minimum value, 0 PPM on the original scale is the same as the median. If we measure a vehicle that has a true RSD HC emission of 0 PPM on the original scale, this would translate into

$$\log(1 + 0) = 0 \qquad (2.6)$$

on the transformed scale. By our rough uncertainty estimate (2.5) we can expect that roughly 50% of the time the measurement after log-transformation would be between -1.06 and 1.06.

After transforming back to the original scale, we get a lower bound of

$$\exp(-1.06) - 1 = -0.652 \tag{2.7}$$

and an upper bound of

$$\exp(1.06) - 1 = 1.88. \tag{2.8}$$

Since a negative value for PPM is nonsense, this means the interval between 0 PPM and 1.88 PPM.

As in Chapter 2.3, we can compare the spread of the within-vehicle residuals with the spread of the data as a whole. The log-transformed RSD HC values have a median of 0. Thus at least 50% of the data are equal to zero, an extreme case of skewed data. Twenty-five percent of the data are above 2.83. Thus, in this case the interquartile range is equal to the difference between the third quartile and the median or 2.83. This is about 1.34 times the interquartile range or the residuals from the linear mixed-effects model. Put differently, the range of the within-vehicle residuals is 74.6% of the interquartile range of the data as a whole.

2.5 Reduced Variation Through Summaries

By taking repeated measurements (RSD90HITs) on a vehicle and computing a summary from them, such as a simple average or the acceleration-adjusted RSD summary, we can obtain RSD values that have a smaller spread than the individual log-transformed RSD Values. (We introduced the RSD summary in Chapter 1.2. We discuss it in more detail in Chapter 2.) It is not possible to see the increased precision (reduced variance or spread) in a plot because we only have one RSD summary per vehicle. Basic statistical theory tells us, however, that the increase in precision is approximately proportional to the square root of the number of RSD90HITs per vehicle that went into the summaries.

RSD summaries made from multiple RSD90HITs are more reliable than single hits. Increasing the number of RSD90HITs, however, even to 8 RSD90HITs does not enable us to predict EC with any certainty. This is in part because the spread of the RSD summary is still fairly wide, even for 8RSD90HITs. In part it is because of variation between vehicles, as we shall see in Chapter 3.

2.5.1 Reduced Variation through RSD CO Summaries

The uncertainty in RSD CO summaries computed from vehicles with 3 RSD90HITs will be plus or minus approximate1y

$$\frac{\pm 0.507}{\sqrt{3}} \tag{2.9}$$

or

$$\pm 0.293 \tag{2.10}$$

For 8 RSD90HITs, by a similar argument, it will be about

$$\frac{\pm 0.507}{\sqrt{8}} \tag{2.11}$$

or

$$\pm 0.179. \tag{2.12}$$

Going through the same procedure as in Chapter 2.3, we find that if a vehicle's true RSD CO value is at the median 0.21% and we have 3 RSD90HITs for this vehicle, then about 50% of the time the RSD CO summary for this vehicle will be between –1.81% and -1.22%. This corresponds to estimated values on the raw scale that lie between 0.154% and 0.285%, 50% of the time.

Similarly, for vehicles with 8 RSD90HITs, we could expect that about 50% of the time the RSD CO summary would lie between -1.69% and –1.33%. This would correspond to estimated values on the raw scale lying between 0.174% and 0.253% 50% of the time.

2.5.2 Reduced Variation through RSD HC Summaries

The corresponding computations for HC are as follows. For 3 RSD90HITs we have an uncertainty of

$$\frac{\pm 1.06}{\sqrt{3}} = \pm 0.61 \tag{2.13}$$

leading to estimated values lying 50% of the time between

$$\exp(0 - 0.61) - 1 = -0.46) \tag{2.14}$$

and

$$\exp(0 + 0.61) - 1 = 0.84. \tag{2.15}$$

Again, we round the lower bound up to 0 so that our interval is [0. 0.84] PPM.

For 8 RSD90HITs, it would be

$$\frac{\pm 1.06}{\sqrt{8}} = \pm 0.37 \tag{2.16}$$

leading, to estimated values lying 50% of the time between

$$\max(\exp(0 - 0.37) - 1.0) = \max(-0.31. 0) = 0\text{PPM} \qquad (2.17)$$

and

$$\exp(0 + 0.37) - 1 = 0.45\text{PPM} \qquad (2.18)$$

Although the RSD HC data, being PPM, are always integer-valued, RSD summaries will not in general be integer-valued. We discuss this in Chapter B.6. 1. Thus, it does make sense to give fractional values for the third quartile of RSD HC summaries.

Chapter 3

Between-Vehicle Variation

Although we see in Chapter 2.5 that we can obtain a more reliable RSD value by summarizing repeated RSD hits on individual vehicles. increasing RSD hits, does not solve the problem of lack of correlation between RSD and Emission Check.

3.1 A Priori Argument for the Importance of Between-Vehicle Variation

Before showing evidence in the data to support this chapter's opening statement, we should point out that we would be very surprised if it were not true. We mentioned this fact in Chapter 1.4. We have no reason to believe that EC was measured without error. That is, if the same vehicle was measured at the EC station several times in one day, we would expect that the measurements would differ from each other. 'They would have positive variance or spread. Yet we only have one such measurement for each vehicle, so we have no way of assessing the amount of spread in EC. (Actually, the raw data with which we started did have multiple EC values for three or four vehicles. This is not enough for us to estimate EC variance.)

Thus, even if we had an RSD value that was extremely reliable, we still would have EC variance keeping us from getting a correlation close to 1 between RSD and EC.

3.2 Evidence for the Importance of Between-Vehicle Variation

In addition to the theoretical argument, we can see this in the data. One way to see this is to make individual scatterplots for increasing values of RSD90HIT.

These plots are made the same way as Figure 1.1, but with one plot for each Value of RSD90HIT. (Since the correlation between RSD and EC is no better for HC than for CO, we only did this for CO.) For each plot, we can compute the correlation. If the only thing keeping us from predicting EC from RSD is unreliable RSD values, then the correlation should go up noticeably as RSD90HIT goes up from plot to plot.

In addition, the standard deviation of the residuals from regressing RSD summaries on EC should go down in a specific, predictable way: It should be approximately proportional to the inverse of the square root of RSD90HIT.

We display the plot for RSD90HIT=3 and RSD90HIT=8 in Figure 3.1. In these two plots, drawn on the same scale, one can see graphically the failure of the correlation to increase considerably from RSD90HIT=3 to RSD9OH1T=8. At RSD90HIT=8 the vertical scatter at higher values of EC is narrower than at RSD90HIT=3. This is not surprising, however, since the plot represents many fewer vehicles. In spite of the somewhat narrower scatter, the correlation is still only 0.42.

In Figure 3.2, we display the standard deviations of the residuals for all values of RSD90HIT from 3 to 8. In addition, we display the "theoretical" standard deviations for RSD90HIT=4 to 8. Let σ_3 be the actual observed standard deviation for RSD90H1T=3, and let σ_4, $\sigma_{5,}$, σ_6, σ_7, σ_8, be the standard deviations predicted by theory discussed immediately above. We compute σ_1 σ_8 as follows:

$$\sigma RSD90H1T = \sigma_3 \sqrt{\frac{3}{RSD90HIT}} \qquad (3.1)$$

In the plot one can see that the observed standard deviations are always greater than the theoretical standard deviations. In the table below, we display the numbers in text form. For RSD90HIT> 3, the only observed standard deviation that is markedly smaller than the value for RSDOOHIT= 3 is the Value for RSD90HIT= 8. For RSD90HIT= 3…….7, the observed standard deviation remains nearly equal to or greater than the observed standard deviation at RSD90HIT= 3. The fact that the observed standard deviation is not proportional to the inverse of the square root of RSD90HIT is due to variation between vehicles.

In the following table, we show in text form the figures that we displayed, in part, in Figure 3.2 and in Figure 3.1.

Figure 3.1: These two plots, on the same scale, show little difference in the width of scatter around the regression line. Although the vehicles with 8 RSD90HITS show greater correlation between CO and RSD summary than the vehicles with RSD90HITS, the scatter for 8 RSD90HITs is still very wide.

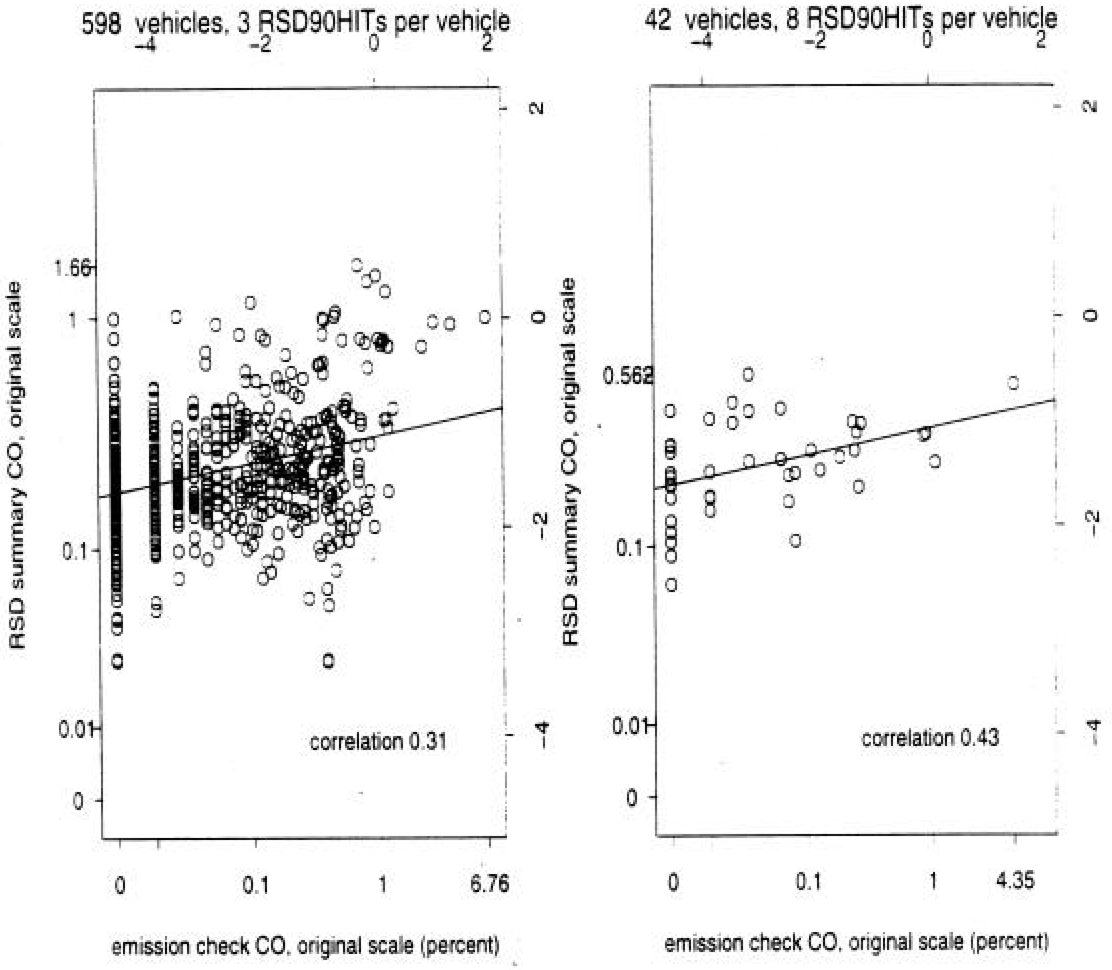
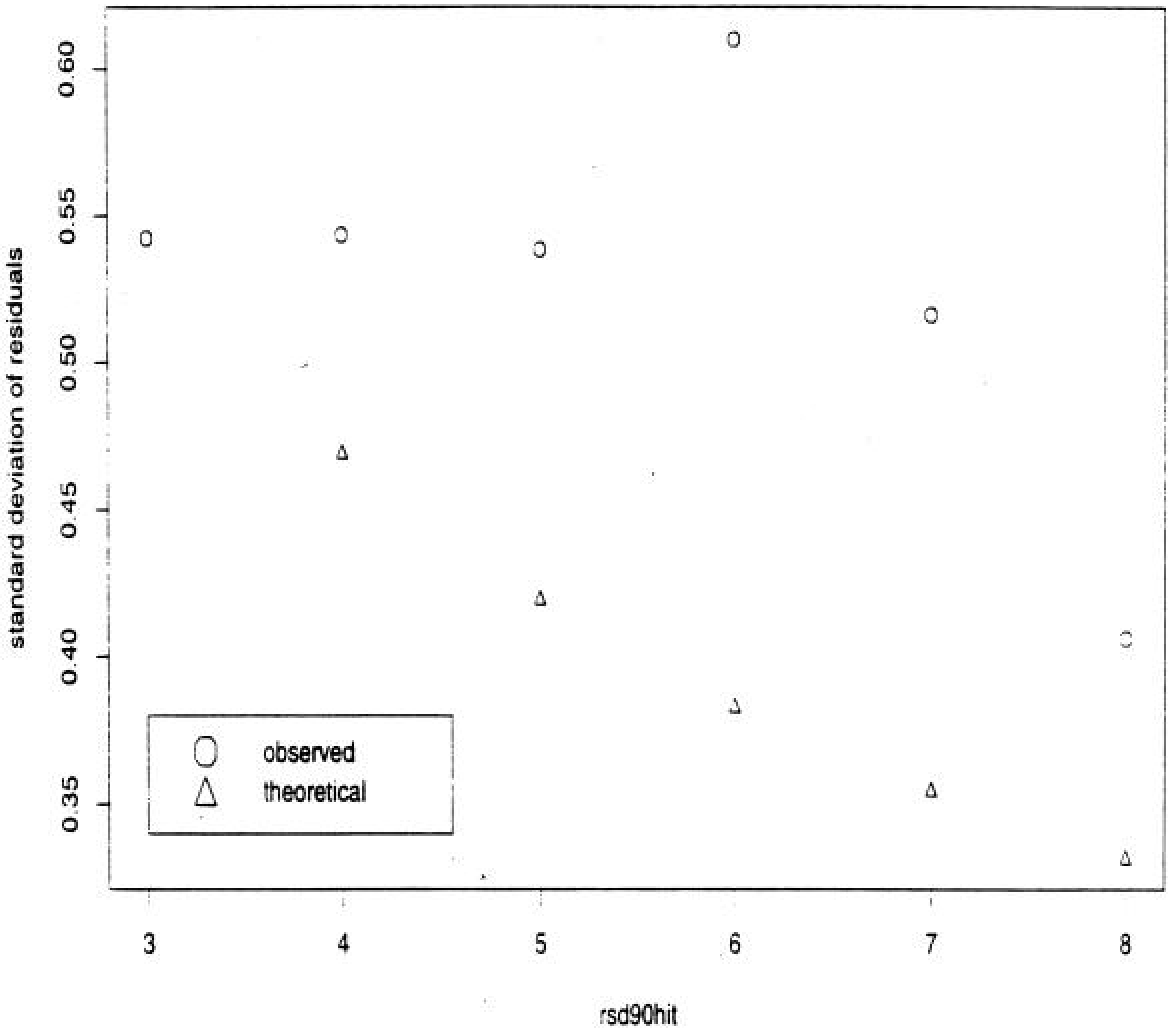


Figure 3.2: Standard deviation of residuals from regressing RSD CO summaries on EC, as a function of RSD90HIT, (We explain the RSD90HIT variable in Chapter B.4.) A separate linear model was fit for each value of RSD90H1T from 3 to 8. On the x-axis was EC: on the y-axis, RSD CO summary. The standard deviation of the residuals from this model was computed and plotted as the observed value. The theoretical values for RSD90HIT=4 to 8 represent the way we would expect the observed standard deviations to go down as a function of RSD90HIT, relative to the value observed for RSD90HIT=3, if the scatter in RSD CO summaries around the linear regression line was completely due to within-vehicle variation.



RSD90HITS	Number of Vehicles	Correlation Between RSD Summary and EC	Standard Deviation of Residuals (Observed)	Standard Deviation of Residuals (Theoretical)
3	598	0.314	0.542	
4	300	0.352	0.544	0.47
5	157	0.399	0.539	0.42
6	93	0.411	0.61	0.384
7	66	0.248	0.516	0.355
8	42	0.425	0.406	0.332

Although a correlation of 0.42 is better than 0.31, it is not nearly enough for us to be able to estimate EC from RSD with any useful degree of precision.

Chapter 4

RSD Summaries

Since the remote sensing dataset contained many vehicles that were measured three or more times, it made sense to compute a summary statistic for each vehicle, such as a mean or a median. Such statistics are more reliable (have smaller variance) than single measurements.

Doug Brown of the Department of Ecology, suggested means, or trimmed means. We did compute both within-vehicle means and within-vehicle medians for purposes of comparison. (The median is the most extreme case of a trimmed mean, “trimming” all but the “middle value.”) We report on this In Chapter C.5. We chose a more complicated method, however, because it enabled us to adjust the resulting summary for acceleration. We call the output of tills method “RSDCO summaries" and “RSD HC” summaries," or simply “RSD summaries.”

Just like the within vehicle-mean or median, there is one summary for each vehicle. The RSD summaries may be thought of as acceleration-adjusted means.

The method by which we computed the RSD summaries involved a linear regression model that is permitted to vary from one vehicle to another. The technical term for this is a “linear mixed-effects model." We discuss some of its more technical details in Chapter C. 4. A reference for this method is [8]. In this report and on the plots, we call the summaries computed by this method “CO RSD corrected for delta" and “HC RSD corrected for delta," or simply “RSD summaries.

The estimate of acceleration in the RSD dataset is called DELTA, as it is in the Hughes dataset. (For a description of how the Hughes people computed DELTA see [10].) Our first step in adjusting for DELTA was to create a new variable “entrdDelta” (centered DELTA”), by computing the mean of all DELTAs (approximately 0.875) and subtracting this from DELTA. The result of this centering was to make the RSD summary easier to interpret. Computed from centered DELTA, the RSD summary represents the value of the log-transformed RSD emission variable after adjustment to the mean DELTA value (0.875).

Please consider Figure 4.1. In this figure, (centered DELTA and log-transformed RSD CO are displayed for one vehicle, LPLATE OOODIT. This vehicle had 7 RSD hits, so 7 plotting symbols appear on the plot. This represents the relationship between centered DELTA and log-transformed CO for this particular vehicle.

The vertical dotted line runs though

centered DELTA = 0

The intersection of the vertical dotted line and the slanted solid line defines the RSD summary for this vehicle. I have drawn a horizontal dotted line through this point so that the RSD summary (approximately – 1.006) can be read off the y-axis.

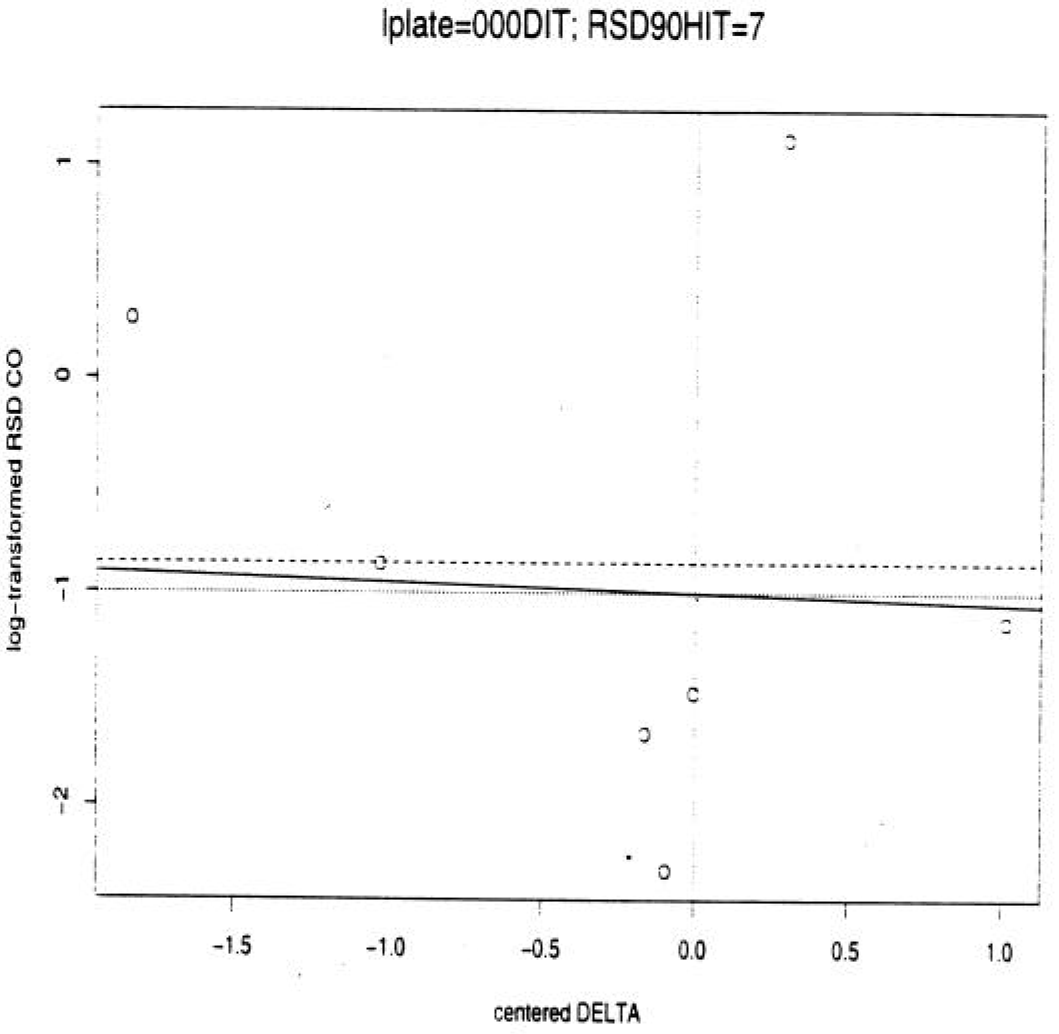
Of course, it was not necessary to actually draw a picture for each vehicle in our dataset to compute RSD summaries. This computation was done automatically by software. I have drawn this picture to demonstrate it graphically.

The horizontal broken line passing through –0.86 is the simple mean of the 7 log-transformed CO measurements for this vehicle. If there were no variations in DELTA, or no relationship between emissions and DELTA, the RSD summary would be very nearly equal to this mean. The fact that the two values are as close as they are relative to the spread between the 7 RSD hits, suggests that the mean would not have been a bad summary. There were differences in DELTA, however, and there was a relationship between emissions and DELTA. The mean would not have allowed us to adjust for this relationship.

We further compare the RSD summary with the within-vehicle mean in Chapter C.5. We consider the within-vehicle median in the same section.

In addition to adjusting for DELTA, it is possible to use the linear mixed-effects model to adjust for weather conditions and RSD location. These adjustments had no effect on the substantive conclusions. Thus for the sake of simplicity we do not include weather variables or RSD location in the model reported here.

Figure 4.1: How the RSD summary was computed for a vehicle. This particular vehicle (LPLATE 000DIT) had 7 RSD90HITs. These are plotted, centered DELTA on the x-axis log-transformed CO on the y-axis. The slope for this particular vehicle, representing the linear relationship between acceleration (centered DELTA) and log-transformed CO, is plotted as a slanted, solid line. This slope was computed by the linear mixed-effects model. The intersection between the slanted line and the vertical dotted line through *cntrdDelta*=0 defines the RSD summary for this vehicle. A horizontal dotted fine marks this value, enabling the reader to read the value off the y-axis. (The RSD CO summary for this vehicle is -1.003.) The horizontal broken line slightly higher represents the simple mean of the 7 log-transformed CO measurements for this vehicle.



Chapter 5

Inverse Regression

In this chapter we give a geometric, graphical approach to inverse regression. For a more technical discussion, we recommend Neter, Wasserman and Kutner [5], Seber [7] and Williams [6] also deal with this subject.

Figures 1.1 and 1.2 have Emission Check on the horizontal axis, remote sensing data on the vertical axis. This kind of display would be consistent with the goal of predicting RSD from EC. Our goal, however, is the inverse: to predict Emission Check from remote sensing data. To do this we need to use a method, which in the statistical literature is called calibration, or **inverse regression.** In this chapter we explain the concept of inverse regression, both geometrically and algebraically. We show how inverse regression breaks down when applied to this dataset.

5.1 Why Inverse Regression?

One might ask why we need to go through the process of inverse regression at all. Why not put remote sensing data on the horizontal axis. Emission Check on the vertical, and predict the latter from the former using straightforward linear regression?

The reason we don't do that: Whereas RSD HC summaries deviate from the assumption of normality (i.e. the residuals from regressing RSD HC summary on emission check are not normally distributed), the emission check values deviate much worse. This goes for both CO and HC. Recall that both CO and HC are discrete, that the lowest value after transformation, corresponding to 0, is arbitrary, and that a lot of vehicles have emission check readings of 0. (Of the 1420 vehicles that we consider in this report, 421 have zero emission check CO and 74 have zero emission check HC.) RSD summaries on the other hand are not nearly as discrete as EC, as we note in Chapter B.6.1. Thus we are stuck with the inverse prediction framework.

It should be noted that the impossibility of predicting EC from RSD in a meaningful way is not a function of the inverse prediction framework. If EC variables did not deviate badly from normality, the insufficient correlation between EC and RSD would still make precise prediction of EC from RSD impossible.

5.2 What We Mean By Inverse Regression

We took Emission Check as the gold standard, remote sensing data as an approximation. Our goal was to find a way, given a remote sensing measurement for a vehicle, or a summary of a set of remote sensing measurements, to predict what the Emission Check measurement would be. As part of this, we had to determine how reliable such a prediction would be. This meant computing a prediction interval, or a range, for that vehicle's Emission Check value. When we say that meaningful prediction is impossible with these data, we mean that the prediction intervals either cannot be computed, or are so wide that they are useless.

In Chapter 5.4 we describe the process of inverse prediction, first in geometric terms and then in algebraic terms.

5.3 Prediction of a New Observation in the Regression Context

Before inverting the process of prediction, however, we must know what straightforward prediction is in the regression context. (This is discussed in [5], pages 76-82.) Consider, for instance, Figure 5.1. Suppose you know that the EC CO value for a vehicle is 0.1%, and wish to predict with, say, 95% certainty what the RSD summary value will be if you get 3) RSD90HITs. The solid slanted line through the middle of the data in Figure 5.1 gives the point estimate. We have drawn a vertical line through 0.1%: follow it up to the solid slanted line
(the linear regression line); note the point where the two lines intersect: draw a horizontal line through that point: read off the point estimate of RSD CO summary from the y-axis. In this case, the predicted RSD CO summary is –1.40 on the transformed scale, or 0.24% on the original scale.

But we want to know not only a point estimate (a single number), but some estimate of how certain we can be of this number. This is what the two broken lines are for, above and below the solid line. (They are actually not straight lines: they only look straight in this example.) By noting where the vertical line intersects the tipper and lower bounds of the prediction band and reading the values off the y-axis, we get a 95% prediction interval for a new RSD CO summary observation on this particular vehicle. We can expect that the RSD CO summary of a new vehicle (new to us, not necessarily a late model vehicle) which happens to have an EC CO of 0.1% will fall inside this interval 95% of the time. Without bothering to compute the interval, however, we can see from the plot that it includes most of the range of RSD CO summaries.

This is the process that we need to invert-do backwards-to try to obtain an estimate of EC from RSD.

5.4 Inverse Prediction: a 95 % Prediction Interval for EC CO

Suppose we have a set of' RSD measurements for a new vehicle, one for which we do not have emission check measurements. We would like to be able to predict the emission check values for this vehicle. Getting a predicted value presents no difficulty. The difficulty arises when we estimate the reliability, or variability, of our estimate.

In Figure 5.1, we may see this geometrically. This plot displays a subset of the data displayed in Figure 1.1-the vehicles with exactly 3 RSD90HITs. Suppose we have 3 RSD CO measurements on a new vehicle, and the summary value (RSD CO corrected for delta) is -1.40.

Geometrically speaking, we get the predicted value by drawing a horizontal line at y = -1.40, seeing where it intersects the prediction line (the solid sloped line), drawing a vertical line through this point, and reading off the x value where this line intersects the x-axis. You can see this on the plot. The number on the x-axis on the transformed scale is -2.21, which happens to be equal to log (0.01 + 0.1). Thus it corresponds to the value of 0.1% CO on the original scale. This is far below the cutpoint (1.2%) for vehicles ≤8500 GVW, 1981 or later.

We can follow this algebraically. We have a regression formula.

$$y = a + b * r + \hat{I} \tag{5.1}$$

where $\in$ represents random error, so from this we get

$$x = \frac{y - a}{b} \tag{5.2}$$

The parameters a and b are the intercept and slope, computed by the modeling function: in this case they are –1.147 and 0.111 respectively. From this we get r = -2.21. Reversing the log transformation, we get predicted emission check CO = -0.01 + exp(-2.21) = 0.1%.

To get an inverse prediction region, we again use the horizontal line which passes through our y (summary RSD) value. The lower limit for predicted x is where the horizontal line intersects the upper limit of the prediction band: the upper limit is where the horizontal line intersects the lower limit of the prediction band.

(The order of "lower" and "upper" in the above sentence is a function of the fact that the regression line slopes up. If the linear regression line sloped down rather than up, one set of "upper"s and "lower"s would have to be switched in the above sentence. It would have to read, "The lower limit for predicted x is where the horizontal line intersects the lower limit of the prediction band," etc.)

But now look at the plot. Because the prediction bands are so wide, both the lower and upper limits are off the scale. This means that the prediction region has lower limit of 0%, upper limit far above the highest CO percent that we can read off the x-axis, which happens to be 3.63 % and far above the cutpoint.

Thus we know nothing more about what value we can expect for emission check CO than if we hadn't measured the car at all. Our exercise has told us that ninety-five percent of the time it will be within the range of the data -a tautology. We already know that it will be within the range of the data 100% of the time.

I chose CO, and this particular number of RSD hits, merely as an example. The same reasoning applies to any of the plots, since they all have prediction bands that are approximately the same width.

5.5 80% and 50% Prediction Intervals for EC CO

In Figure 5.2 we try narrowing the prediction band by settling for 80%, or .50% certainty. Following the same procedure as in Chapter 5.4.we find that even with 50% bands, the prediction interval for EC on a vehicle with RSD CO summary of -1.40 includes the entire range of the data.

5.6 50% Prediction Interval for EC HC

Let's follow the same procedure for Emission Check HC as we have just followed for CO in Chapter 5.1. Please refer to Figure 5.3.

Suppose our RSD summary for a hypothetical vehicle, on which we do not have an Emission Check HC measurement, is 1.2295. The parameters of the linear model in this case are a = 0.866.b = 0.152. Our predicted EC HC on the transformed scale is

$$\frac{1.2295 - 0.866}{0.152} = 2.3976$$

This translates into

$$(\exp 2.3976) - 1 = 10\,ppm$$

on the original scale.

The horizontal line intercepts the upper boundary just below EC=0 ppm on the original scale. It intercepts the lower boundary at a point between 100 ppm and 217 ppm. Thus if we can take this prediction interval seriously, we

Figure 5.1: Attempt to generate a 95% prediction interval for a new RSD summary observation. Data plotted are vehicles with RSD90HITs=3. Solid slanted line is linear fit to the data. Slanted broken lines are the borders of the 95% prediction band for new RSD observations, given Emission Check (usual, not inverse regression). Horizontal broken line represents a hypothetical new RSD summary for a vehicle for which we have no EC value. Vertical line marks where
it intercepts the prediction line, and where vertical line intercepts the x-axis we can read off the predicted EC value for the hypothetical vehicle.

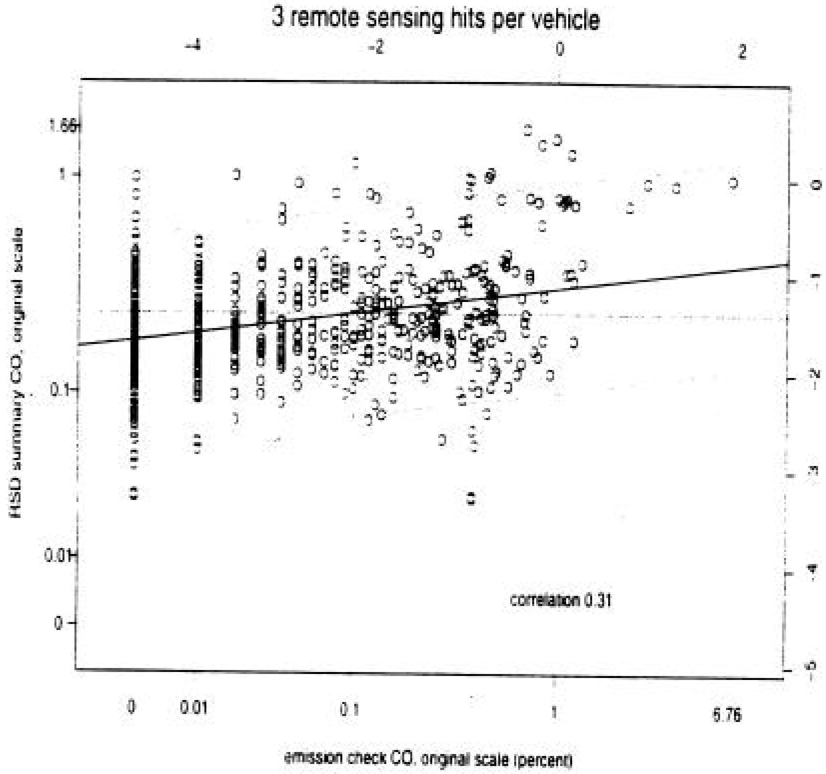


Figure 5.2: Figure 5.1 revisited: Now we are settling for an 80% or 50% prediction interval. Same data as in Figure 5.1. Outer two slanted broken lines are the borders of the 80% prediction band for new RSD observations, given Emission Check. Inner two slanted lines are 50%.

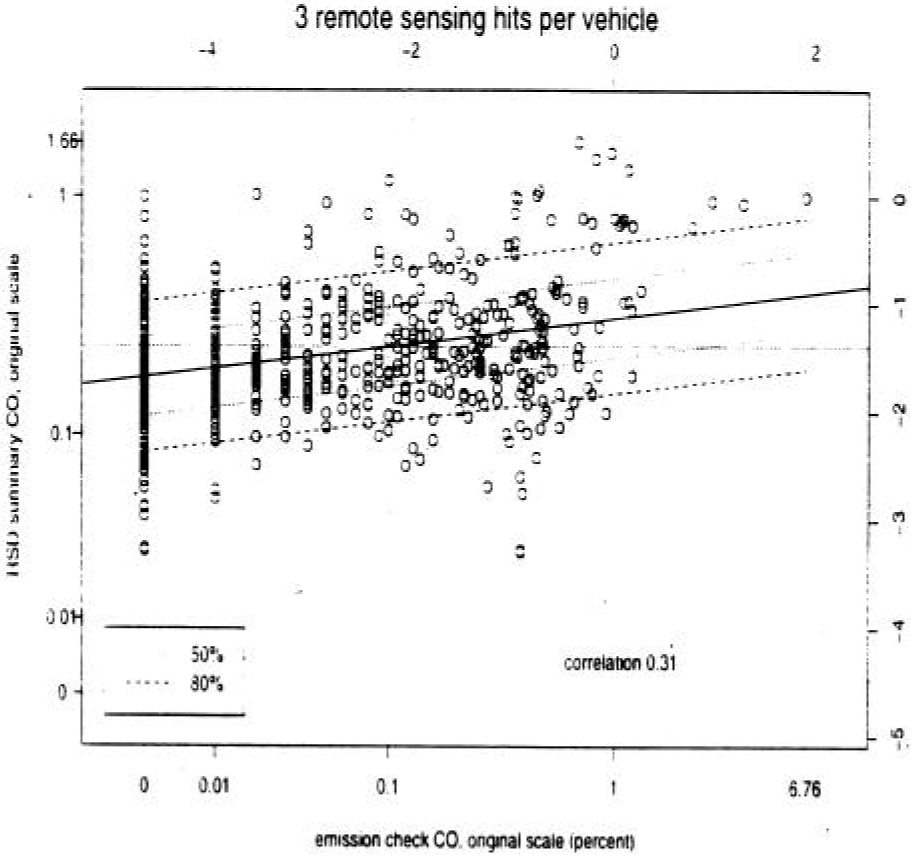
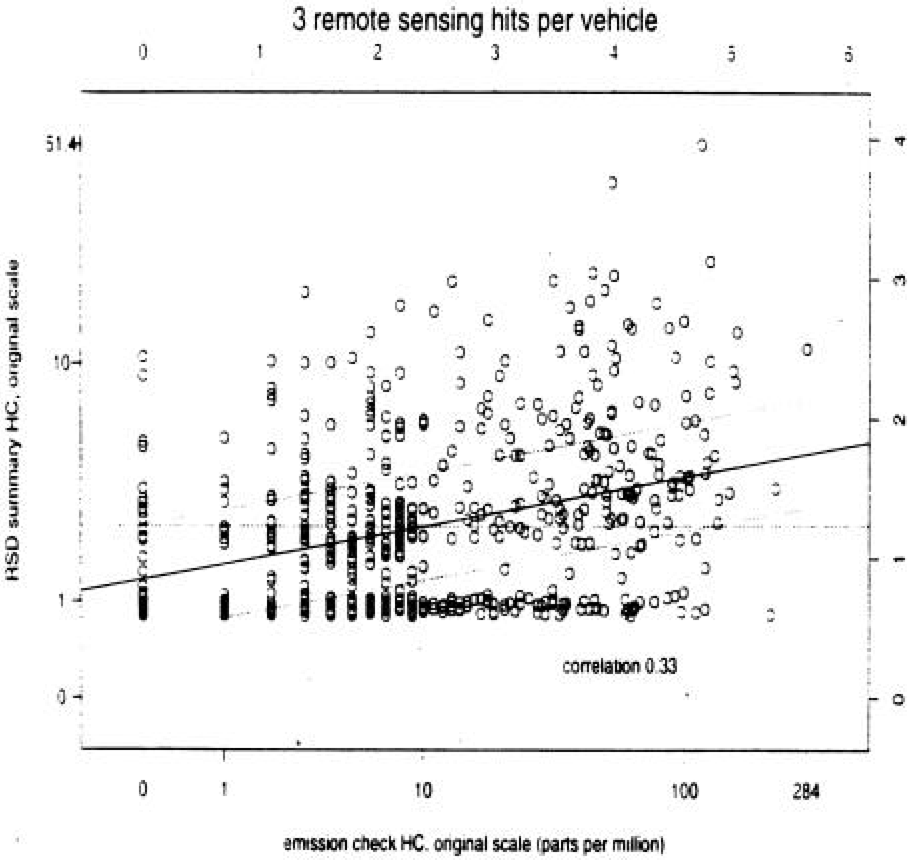


Figure 5.3: Inverse regression framework for predicting a new Emission Check HC value from an RSD summary HC value.



can expect that 50% of the time, when a vehicle has this value for RSD HC summary, the Emission Check HC value for this automobile will be below the 220 ppm cutpoint. This doesn't come as a surprise, however, because every single EC observation in this particular dataset is below the 220 ppm cutpoint.

5.7 Differences Between CO and HC Plots

The HC plots look different from the CO plots. The RSD summaries are clustered toward low values, with a few high values. In other words, they are skewed towards high values. As a result, the lower prediction bound is actually off the scale in many places. In Chapter B-6.2 we mentioned the fact that HC is more skewed than CO. We deal with this further in Chapter C.3. The fact that the log transformation of HC does not succeed in making HC residuals close to the normal distribution, means that we must interpret conclusions based on normal theory with caution. For instance, we should interpret the 50% prediction intervals for EC HC that we have been computing, with caution.

Appendix 3: RSD Activity in other states and provinces

This summary was current as of the end of May, 1998. All RSD programs use infrared sensing unless otherwise noted.

Arizona: Previously, six RSD units from SBRC were in use in the Phoenix area during 1995 and 1996. Clean screening and gross polluter identification was being studied. It was determined that clean screening was unreliable and that there was no way to verify false passes. A gross polluter ID program was designed and put out for bid. Envirotest initially expressed interest in bidding for the contract but withdrew after expressing their disapproval of gross polluter identification. Envirotest expressed a preference for clean screening. Arizona is looking for an alternative contractor and is considering laser technology.

California: The BAR had been using twenty RSD units and forty were planned by 1999. Primary goal was gross polluter identification. The program was cancelled in March 1998. Envirotest expressed a preference for clean screening and quality control issues were in question. RSD is still a possibility. Other contractors and technology are being studied.

Colorado: An RSD study was conducted in the City of Greeley. 1,500 vehicles were tested through May 1997 by RSD daily during ideal weather conditions. It is planned that a clean screening program will begin in basic I/M areas in January 1999. The plan is to clean screen 50% of the fleet with two RSD readings. The study indicated that the repeatability of RSD was questionable. It was determined that if a big enough portion of the fleet is tested it will work out to only missing about 6% of potential CO failures with a cut point of .5%. The approach is to pass cars not fail them. A clean screened vehicle owner will receive a bill for $15.00 from the contractor. The owner may still drive through a test station.

Florida: A supplemental roadside RSD gross polluter program has been recommended as a deterrent against tampering.

Idaho: Proposed: Clean screen 35% of fleet. Identify 4% of fleet as dirty. Remainder of fleet to test as usual. Pay fee at registration. Use plate recognition to track out of area travel. The study indicated that statistically RSD does not work. Logically it should, therefore RSD will be part of the plan. A static cut point should not be used; the goal is to consider the cleanest 35% of the vehicles as clean. A new emission inventory indicates that NOx testing will be required by 2000.

Massachusetts: Plans to set up a study to determine RSD accuracy for a clean screen program in Berkshire County.

Missouri / Kansas: A study had been planned to look at gross polluter identification but was dropped. Re-formulated fuel will be used in place of an I/M program. Radian will be hired to study the effect of the fuel on the fleet.

New Jersey: RSD units were placed at test station entrances for comparison studies. Future RSD studies are planned. ASM testing is scheduled by 1999

New York: Envirotest awarded contract to clean screen vehicles from testing. The fee is billed by the contractor to the vehicle owner after being clean screened.

Ontario (Toronto): Is experimenting with a RSD unit located at the test station entrance to screen out the low-emitting vehicles and send them home without having to stop at the test center.

Texas: RSD was to be used to target high-emitting vehicles commuting from adjacent ozone non-attainment counties. RSD was also to be used as a program validation tool to test vehicles exiting from test and repair sites. SBRC was to be the contractor for this project. In August 1997 Envirotest declined to bid on the project unless the gross polluter identification was dropped for clean screening. The TNRC searched for another contractor and awarded the project to Tracor. Tracor will use laser beam technology and is scheduled to begin testing September 1998.

Virginia: Virginia has a decentralized ASM program with gas cap checks. The state has set aside plans for an RSD program. An RSD study found low correlation with tailpipe inspections and raised concerns about the effect of unknown and variable engine loads on each vehicle's on-road emissions performance.